Medicaid Innovation Accelerator Program
Webinar on Missing Data
June 7, 2018

Hannah Dorr: [intro logistics]

Jessie Parker, GTL and Analyst on Medicaid IAP Data Analytic Team, Data and Systems Group, CMCS: I'll start by providing a quick overview of the Medicaid IAP program. First, we'll learn about the different types and patterns of missing data, as well as the problems they can cause when trying to make conclusions based on data with high rates of missing data.

Next we'll hear about Alabama Medicaid's experience with missing data and what they do to address this problem.

Lastly we'll end with a joint Q&A session where you're encouraged to submit questions via the chat box for either speaker. You can submit them at any time.

Our first speaker will be Thomas Flottemesch, Senior Research Leader, IBM Watson Health. Dr. Flottemesch is an economist and statistician with experience in the areas of clinical design, utilization, and quality improvement. He previously partnered with Minnesota's Medicaid program to evaluate clinical implementation, patient adherence and the impact of a recent oral health initiative. Dr. Flottemesch will be presenting on the types, patterns, and impacts of missing data.

Our next speakers are from the Alabama Medicaid agency. Chris McInnish is the Director of Quality Analytics and Drew Nelson, MPH, is the Director of the Quality Assurance Division. They will both be speaking to their experience addressing missingness in their maternity data by pulling on a variety of data sources to increase accuracy and completeness.

This webinar is produced through the IAP. We are in the data analytics area, and one of the approaches we use to increase data link capacity within our state Medicaid agencies is to host webinars such as this one on various data-related topics and challenges that are relevant across a wide array of states.

Our primary goal for today's webinar is for you to gain an appreciation for the seriousness of missing data problems and thinking about why and how you can address missing data in your own state environments. First we'll cover the different types of missing data and the issues they can cause, before beginning to cover ways in which you can address these problems by creating a more complete data set. There are also analytic solutions to missing data, which we'll cover in-depth in a future webinar.

Our first speaker is Dr. Flottemesch with an overview of what missing data is and why you should be concerned about it.

Thomas Flottemesch (TF): First I'd like to give a broad overview of what we're talking about when we're talking about missing data and put it in the context of statistical analysis. Then I'll walk us through how missing data may appear within a data set and provide some examples and real world examples of how this can actually influence the conclusions we draw from data.

The issue of missing data, what it really refers to is we're talking about observations that are missing certain types of variables, so they're incomplete. What we're not talking about are observations with no

data or things that we can't observe. There we're getting into more of an empirical issue of unobserved variable bias. What we're talking about is we try to collect data but there were just certain variables we weren't able to collect for whatever reason.

With missing data what we really need to know are two things: Are the missing data informative about the data we observe? In other words, what we see, is it impacted by what we're not able to see? And can we ignore the missing data, because if we do, will it lead to incorrect inference or conclusions? I like to refer to missingness, or the issue of missing data, within the broader context of statistical analysis. When we do a statistical analysis what we're really trying to do is understanding the data-generating process. How are the data being generated? Given that data-generated process, we ask questions of the data. We ask what happened, what did we observe, and how is it happening, and whether or not if we have two groups did it happen differently between the two groups, or if we're forecasting, will it happen again.

The issue of missing data within a data set really goes to that probabilistic inference. We need to know two things: What is the pattern? Which values are missing within our data set? And what is the mechanism? How is what we cannot see, how are the missing data related to the observed data? Are they just at random? Are they completely at random? Or are they part of the underlying data-generating process that we need to read into our model?

If we think about the structure of missing data, the simplest type of missing data is really univariate missing data. That's when there is a single variable and it is missing data. This is most critical if we think about being unable to observe our outcome or being unable to observe the factor of interest. When I talk about factor of interest, I'm really saying what is the key thing we're trying to see if there's a causal factor within the model and we have other variables that we're adjusting for.

More frequently, though what we encounter is multivariate missing data, where within an observation, where we're trying to do, for instance, risk adjustment, there will be multiple variables that are missing data. This pattern of missing is varied across observations and across variables. The key factor we want to ask about or the thing we want to know about is how data are missing, because how they are missing really informs a strategy for addressing this.

The first thing I'd like to talk about is this concept of monotone or nearly monotone data. There's no perfect test for this, but the idea here really speaks to the efficiency of any strategy we implement. The other thing I'd like to point out is within any data set, different variables can have different underlying causes of why they're missing, and we may have to address that differently as well.

If we talk a monotone structure, just a brief minute on this. If we look at a sample data set where all the missing variables have little red question marks, we look within our data set and really have four groups of observations. At the bottom we have group D, which are complete observations. Above that we have C, which is missing one variable, and then groups B and A. The key thing, if we have monotone missing data, is that we can conserve information if we're talking about something, for instance, multiple imputation, or if we're talking about a maximum likelihood strategy, we increase the likelihood of convergence or reduce in a multiple imputation set a number of imputed data sets we need if we leverage this structure of the data.

Specifically we can use our complete observations to fill in the gaps or the almost complete observations of group C. Then we can combine those two groups to fill in for B and then for A. How do we do this? It's

often built in. If you're a SAS user there are PROC multiple imputations. How you screen for it, first I would use PROC 3 just to look at how many missing data and if there is a pattern. In R, the MICE Library, multiple imputation with chained equations, has this built in. If you're a Stata user I believe it's also built into their MI procedure. If your data doesn't quite meet these requirements, check your log file in SAS, it would definitely tell you. In R I think it would just stop. I'm not quite sure how Stata would handle it.

If we turn to the mechanism of missing data, it really goes back to 1976 and Rubin developed a typology of missing data mechanisms. The first is missing completely at random. We have missing data that is unrelated to our study variables. In other words, the data and the complete observations we observe are a perfectly unbiased random sample. A rule of thumb, this is generally a bold and pretty much unprovable assumption. The safest route is usually to go to a missing at random assumption and whether or not the data are missing does not depend on the values of the missing data. In other words there's an ignorability here. We need to address or populate these missing data, but we don't need to directly model why they are missing. And the data we observe can predict the data we cannot.

The final situation is where things get the most complex. This is when data are not missing at random or missing not at random. Whether or not data are missing depends on the values of the missing data. We have to directly model why the data are missing; what's the probability of observing something? That means we have to add a layer of complexity to our final analytic model rather than doing analysis on an imputed data set.

It's better if we walk through a real world example. Let's consider predicting annual healthcare costs. We will adjust for items—for instance, family history, cancer, diabetes, depression scores/PHQ-9, and smoking status. We collect these data, pull them from electronic medical records or billing data, and we see the following pattern. We see that we have all the costs from our utilization data, but as we can see with the red question marks, different observations and different groups of observations are missing different types of data.

So the question is what can we do about it? Why are all these data missing? How do we address the issue? If we look at our analysis, family history—people may simply not know whether or not their grandma did or did not have cancer or what type of cancer they had. In other words, we could move into the framework of saying those family history data are incomplete because they're probably missing completely at random. In contract, depression scores—men may be far less likely or willing to respond to a PHQ-9, so we can look at these data as missing at random because even though they're not completely at random, because we know there's been an association between men and whether or not they completed a PHQ-9, we can't necessarily say men are more likely to be depressed than women. Then finally, smokers—they may just not want to share their smoking status. In other words, whether or not we observe smoking status is probably directly linked to whether or not someone's a smoker. That means it is missing not at random. We have to model the likelihood someone's a smoker and respond to the survey.

Missingness mechanisms are ignorable in the first two cases. In the final case, we have to model it directly, either some sort of pattern or mixture model. So for an illustration of a real world example of missing completely at random, for example I've run across these maternity episodes. The task is we wanted to develop risk-adjusted estimates of maternity episodes, but we only had one year of data. We had risk adjustors for maternal health, gestational age, prenatal care initiations, but we had a missing data issue. These episodes are available only for the start of the year. Remember we had one year of data, January through December. However, many of the pregnancies started in the prior year. So why are the data

missing completely at random? When a pregnancy occurs it's probably independent from the healthcare or the complexity of that episode. In other words, the missing data are not related to the study variables.

So what strategy can we use? We could do a complete case analysis and use only the variables we observed, but then we see this pattern. We'll have full data capture if you got pregnant at the beginning of the year. We'll have partial data capture if you got pregnant in the second quarter of the year, but they will be confounding with complexity. And if you got pregnant in the last half of the year, we're only going to see a complete episode if it were a premature birth or a very complex case.

What's our other strategy? A month-by-month imputation. When I say imputation I'm not talking about multiple imputations here. This is an example of a maximum likelihood type imputation where in the first step we modeled the expected utilization for each episode. Then we used the model to impute missing values for the unobserved months and reassembled the episode. It looks something like this. We saw certain months, and for other episodes we may not have observed months 6, 7, 8, 9 and 10. But using data from the complete episodes, we were able to populate or impute the missing months and then reassemble the episode.

How did it impact our analysis? What we really saw is in the third line of the upper table, the full-term births, if we only use a complete case analysis, we had very few of those, or 45. We basically threw out—comparing at the bottom half of the table—the full-term episodes, if we would have imputed or filled in the missing gaps of those months. That would lead to a different inference if we looked at average costs or this standard deviation of the variability of costs.

Let's look at the next example, men less likely to complete a depression score. This is data that's missing at random. In other words, it's correlated or one or more of the variables in our data set can be used to predict this score. Here's the illustration of an example I came across: missing at random in emergency department utilization. The idea is to estimate the relationship between behavioral health integration and ED use among substance use and mental health patients. Inclusion criteria were Medicaid enrollees with 10 or more months of semi-continuous enrollment. The key point here is 10 or more months.

For missing data issues, we wanted a full year of ED utilization. The issue is that ED utilization follows known cyclical patterns by month. In other words I have an illustration here on the bottom right. You can see that in certain months you're just more likely to go to the ED. So ED visits in unobserved months are predictable by when they are missing. If we would use—and I've done this in the past—a typical strategy, K-means or nearest neighbors, we could, one strategy, populate missing months with the person's average from observed months. That's a very common approach. But that ignores the seasonality we talked about in the prior slide.

Another is to populate missing months with an average from age and gender matched peers. The issue here is we would be deflating the variation and potentially introducing overpower or giving ourselves an ability to have a false positive within our data set.

Our second strategy is multiple imputation with month covariant/fixed effect, and this is directly implementable within SAS, PROC, multiple imputations or within R with the left-centered means. People like to use the multiple imputation change equations in R but it assumes an underlying normality, and ED visits are more of account data, so I would caution against using that strategy. But if we compare the result using just averaging or the K-mean imputation, we can see that in the months of April, June, July, and

September, those impeded values gave us smaller average numbers or predictive means of ED utilization relative to using a seasonally adjusted imputation strategy. So we will have underpredicted the relationship within or during those months in our final analysis.

Finally, smokers tend not to share smoking status. This was our example of missing not at random. We have to model the missingness mechanism and selection or pattern mixture models. The example I have here is something that is actually an ongoing analysis but I thought it would be interesting to illustrate it. It's a maternal episode. Estimate the total amount spent on delivery by gestational age.

The missing data issue is emergency transport cost is tracked differently in different systems and not available in fee for service versus managed care individuals. Why are these data missing not at random? The amount and type of data collected is associated with the payer type. So if my analysis was focusing on comparing cost differentials between managed care and fee for service populations, I would have different data captured across the population, and without directly addressing that I would make incorrect inference. In other words, to go back to Rubin's typology, the pattern of missing data is dependent upon the pattern of missingness.

An example just of raw data, I particularly want to point to the preterm births. Among fee for service we have no emergency transport costs, so the average cost per delivery is approximately $5,000. We compare that to a managed care population, that average cost is approximately $20,000. It would be a bold statement to say that fee for service is that much more efficient at preterm births. What we are missing are one of the key cost drivers of these preterm episodes and that is the emergency transport cost, particularly in rural populations where a higher level or level 3 birth center needs to transport complex deliveries to NICUs or higher level facilities.

What do we need to do or what choices do we have here? We have either excluding all the emergency transport costs, or inserting or imputing some sort of emergency transport costs, or do some sort of Heckman adjustment for a pattern-mixture model to account for this differential data capture.

I've kind of hit you with a fire hose in regard to missing data and apologize if people are a bit overwhelmed. To summarize, the key points I wanted to touch on with examples are why do we care about missing data? And why we care and why we should be concerned really does link back to how the data or why the data is missing? If they're missing completely at random we could do a complete case analysis, but we are lowering our precision. If we have missing at random, it can actually affect our analysis and interpretation, and we should use the data we have to try to address and populate the data we do not.

If the data are missing not at random, there's an underlying mechanism here we need to address. Otherwise we introduce potential bias into our inference and we can come up with the wrong conclusion.

The other thing I really wanted to address is how does missing data appear in data sets, and finally, what can be done about it? There's always ignore a complete case analysis. We can impute, and when I say impute there's a variety of ways—multiple imputation is one of the most popular ways to address it. Maximum likelihood is another variable approach as well. Finally, we can model the missingness directly, particularly when it comes to data that are missing not at random.

JP: Thank you. That was a solid foundation for why we should be concerned about missing data, whether you're collecting data, building a data set, running summary statistics or moving on to more advanced

inferentials. Whenever you're trying to make conclusions based on your data you need to be aware of issues that can be introduced by missingness.

Our next presenters will be Chris McInnish and Drew Nelson from Alabama Medicaid. They will speak to how they address missingness in their data by pulling from multiple data sources to form a more complete picture.

Chris McInnish: Our approach to missing data is different in primarily the problem we're looking at and the various data sources we have. We're going to really talk about our maternity program and the data we use to analyze it. Using the standard diagnosis revenue codes, procedure codes, all those codes to identify deliveries, really results in a high error rate for us. Where we find claims that are just not deliveries and we have duplicate delivery dates for one pregnancy, a couple of quick examples: In 1995 we had a 95-year-old female that I got a crossover claim for on the hospital and a separate crossover claim from a physician showing that she had become a mother. We also had a 1-year-old male who halfway became a mom; ER visits where the mother goes to the ER for morning visits where the diagnosis code is morning sickness with delivery, probably a coding error because a couple months later she had another delivery; and an unlucky woman who had two babies a week apart. What we really think happened was the second delivery was more associated with a readmission for postpartum depression, and it really came from that. Additionally, multiple sources provide inconsistent data for situations such as gestational age, the number of prenatal visits, and weight. So what do we do when we have multiple sources with different gestational ages? How do we handle that?

This variation, in claims data and certain data, is only available in birth certificates and necessitates the link with vital stats, so we pull in the birth certificate data to be attended, but a simple match of birth certificates didn't work. There are a number of issues. One, the birth certificate data and vital stats, where the birth certificate indicated Medicaid delivery, it had a 20% error rate where we had deliveries that we actually paid for that were not showing up with that mark on the birth certificate checked. Additionally, I had problems matching by names, Medicaid paying for deliveries that occurred out of state in a lot of hospitals just across the border, and other things that just resulted in our counts being glow and our numbers not coming in accurately.

What our approach is is to pull together multiple sources of what I call semi-truth. These are different things we're going to look at. We'll pull in hospital claims, physician claims, vital stats. We have a case management system in our maternity program and we're using their chart reviews and their claims. And then we try to build a hierarchy for how to use the best source for each element of a delivery.

First things first, we have to know which is an actual delivery and which is not a delivery. As with all this filling in missing data, nothing is perfect. We start with all possible deliveries from claims data. We run through and find anything that might be a delivery including the 1-year-old male and 95-year-old female, pull all that together, de-duplicate where we find similar dates and similar mothers, so we try to make a de-duplicated list. Then we validate it, ultimately hoping to find both the baby's identification and match the mother's delivery. The case management piece, about 60-70% of our deliveries have the case managed so it's not 100%.

So validation of delivery, this is how we find out if we want to count a delivery as actually valid or if we want to consider the information on that claim to be in error as far as counting it for our delivery. This is

really a process. We go through a series of rules. If a delivery meets a rule we consider it valid. If not, we don't consider it a valid delivery.

So at the start of this process vital stats provides all data on delivery. We give them all the information we have on delivery and they provide those that match back for all our deliveries and for all the children born within a year. As we go through that process, if we can match our claims data to a birth certificate then we consider it truth, and that works. That gets a large chunk of our deliveries approved as valid.

Next, if the hospital that delivered is out of state—we're in Alabama and if they're in Georgia and that's the hospital, we obviously would not have a birth certificate match for that because we only match with Alabama vital stats, so would not consider that true. We then look to see if we have two independent sources. So if we have a claim from a hospital and a claim from a physician or one of our contractors, then we would consider that true. But then we go back and exclude where we have multiple deliveries. So if for some reason we had counted one in vital stats and go back and find one we had multiple sources, we would run that out and also provide a reasonable check for the mother's age. So if the mother is less than 9 or over 60, we exclude those as improbable.

Next we go to the delivery date, as another piece of information that's important in our delivery file is understanding when the delivery date was. In that we have a kind of hierarchy. The most important information for that is vital statistics. If we have a vital stats match we use that match as the official record of the delivery date. From that, if we don't have a vital stats match then we go through a case management claim followed by a physician claim, and in the hospital date of admission is the last piece. In that is an order. The birth certificate is the official record, so we see that as the most valid piece. But if we get to the end and don't find anything, we have the hospital date of admission. My first child, like many, we went to the hospital one day and had our baby early the next morning. So we understand that that's not valid. If we don't have anything else it's the best we have and it's our last resort.

Another piece of information we're looking for is gestational age. For this we also provide another hierarchy to get through the sources of truth in order of vital stats, where we have the vital stats match. If not, we can use a claim from a hospital or physician using diagnosis codes and we also have some other special ways on some of our claims that the hospital and physician can include that. Then we use our case management system last. In all this we do a reasonableness check. For example, we exclude data that says the gestational age was less 18 weeks or over 45 weeks. We have had some that show gestational age as well over 45 weeks.

The next piece is the number of prenatal visits we're looking for. We do a little bit different number from this. Instead of considering one source valid, we look at the multiple sources, vital stats, claims from physicians or case management system. We really look to see which one has the most. The logic is that if the birth certificate says there are only 4 prenatal visits but we're looking here at 14 claims from an ob-gyn saying they were prenatal visits, we believe the claims. So whichever one is the highest volume is the one we consider the most accurate for this specific measure.

Birthweight—and we have a reasonableness check on this—but the source of truth we use vital stats first and the claims from physicians and hospitals followed by the case management system next in order. For postpartum visits we use the claims from physicians followed by the case management system with whether those happened.

There's other information out there. One of the pieces is we also need information on newborns. If you're Medicaid you probably have this same situation. We cover a lot of babies, some of which are quite sick, where we did not cover the pregnancy or delivery. So when we start to look at our infant program, which is often tied to prenatal factors, we want to try to capture that information as best as possible. So we try to grab that information not only for the babies we cover but also the babies we did not cover, the pregnancy portion of their life. We also try to find the claims for deliveries so we can match the infant ID, and we use vital stats and eligibility information to merge this information together.

The final product that comes out of this, what we produce for our analytics, is two files. One is a maternity file with all deliveries. This is one line per baby so one delivery with multiples would have multiple lines, and we manage that together. We use all the different sources to come and find that that's value. So there's one field on that table that shows the gestational age, one field that shows the data delivery, and we've used our process to identify the best source and filled it in at that point. We also have an infant file. This is one line per baby born per year. We've gone back to 2010 pulling this forward. We're looking for any baby that had eligibility within the first 3 months of life, and then pull all the information we came from many different sources for that. So if we pay for delivery we would have physician claims for prenatal visits, all that information. If we only had it in vital stats then we can only get it from vital stats. That just lets us fill in the gaps as we need to.

Drew Nelson: Now that Chris has talked about his and his team's work with analytics, we want to give you a couple examples that we've used on the evaluation and policy side. One of the things we're definitely looking at is infant mortality and historically we've been pretty bad.

This first chart shows that we know we've always been pretty consistent with the number of deliveries, but by bringing in our different sets of data, we were able to go from about 28,000 to 30,000 deliveries per year to a truer number of 35,000. Because we were able to bring in additional information from whether it's a high-risk delivery which falls outside our maternity contractor; our ER and emergency deliveries have not gone through on our normal claims process; third-party payments, whether the child comes in and the hospital might mark it as a Blue Cross/Blue Shield or Medicare baby but it's actually been paid for by Medicaid, we would not normally historically have gotten those. Also we talked about the out-of-state deliveries.

We do think the more numbers the more data we get and the better data we get from vital stats and our other sources actually improves our numbers. Some of the reports we've historically looked at are prenatal care, with more than 50% of delivering mothers having less than the recommended number of prenatal visits, one thing is we are using as many sources as we can to make sure we get the correct number of prenatal visits, whether it comes in from an ob-gyn, our contractors or our birth certificates. We divide our state up into 14 different maternity care districts, and the last two years District 10 has not had a contractor so for all our case management where our core measures rely on those sources, we haven't been able to pull in that data. So having vital stats and claims data as well as some data from the hospitals actually fills in that picture for us so we have a full, statewide analysis to look at.

We also look at prenatal postpartum care. One thing we looked at it is over 30% of our babies had at least a first visit after 18 weeks. Postpartum visits, though, we can't always document for about 70% of the women, for those are the women who do not use our contractors. So by combining looking at those deliveries paid for by the contractor or the noncontractor, we're able to get our full deliveries as well as all the postpartum and prenatal visits, whether that comes from a hospital, an ob, or even vital stats.

Eligibility shows where our mother's eligibility status has been. It's a big issue particularly where the vast majority of our women have historically been SOBRA or full Medicaid, but we are starting to see a bigger change where it's shifting. As well as a lot of the eligibility after 4 months, if we're looking for certain claims or the delivery's an alien or the child's SSI, we wouldn't have gotten that. This shows one thing we're looking at is where our mother's eligibility is after delivery.

For delivery costs, we also look at this to see is there a difference between the two. We're looking at total amount of cost 2 months prior to 10 months post-delivery. It includes all costs incurred by the mother from any provider whether it be an ob or physician.

Infant costs is for all infants covered by Medicaid whether the mother's delivery or prenatal care was covered on Medicaid or not, which is about 38,000 children. The average first-year cost of a NICU is around $52,000 a year compared to the average first-year cost of an infant without a NICU stay, which is less than about $3,000 a year. The ICU costs account for approximately 60% of our total infant costs, and newborn and NICU costs are a major cost driver with significant dollars in long-term NICU stays. Some of the other costs that might be included are professionals, specialists, hematologists and things like that, anything besides inpatient.

One other thing we've been really focused on because of the opioid epidemic is looking at our neonatal abstinence syndrome. We have been consistently seeing an increase on this. Geographically it's concentrated in north-northwest Alabama and Walker County, that really bright purple upper left-hand quadrant, one of every 20 babies is born addicted to opioids so this is one of our biggest issues we've been looking at. We've been partnering a lot with our public health as well as our pharmacy program and PDMP program.

That's just a couple examples of standard dashboards and reports that Chris's team, by using multiple different data sources, has been able to provide to our administrators and policymakers so we really have a better idea.

CM: And this merging data from different sources is what's allowed us to do this analysis fairly quickly and keep it up-to-date so we can let the policymakers know the information they need to know.

JP: Thank you. That was a really good example of how you can adjust missing data problems in a real world setting by referring to multiple sources of information to create a usable data set to help your program analyses. We'll end with a Q&A session. Tracy will facilitate.

Tracy: The first question is for Tom, our first speaker: *How exactly do you determine that data is the MPAR type? Would you also use maximum likelihood estimation?*

TF: There's actually a test for missing completely at random and it's called Little's MCAR Test. I believe it's actually implemented in SPSS, Stata and SAS. I don't know if it's formally implemented in R. It boils down to a pie square test. It would indicate whether or not there appears to be some sort of pattern within the missing data that is correlated with the data you are able to observe.

However, I want to give a word of caution about using this test. Even though I'm trained as an economist and economists love to develop statistical tests for everything, the null hypothesis of Little's test is that the data are missing completely at random, so you're looking for evidence that they're not missing completely at random. However, if you think about a sample size, the smaller your sample, the less power

any statistical test has. But when we think about missing data, the smaller our sample, the greater the cost of mis-specifying the missing data problem is. So these are kind of working in opposite directions.

So you can use this test and it has a large sample size and indicates the data are missing completely at random I think you're perfectly safe. If you have a smaller sample size, I would consider even if you fail to reject the null and Little's test still employing some sort of multiple imputation strategy or maximum likelihood strategy. I guess it was another question about maximum likelihood as far as a strategy for dealing with missing data. I know this will be addressed in a later webinar. The idea between the two with maximum likelihood is what you're really doing is you can think about two different equations within a model. The first equation is what you're really interested in. The second equation is predicting your missing data. So you're converging to what is the most likely value for that missing data given the data I have, and then what the inference is.

The biggest benefit for using that strategy is that you have a deterministic result. In other words, you get one result. When you use multiple imputation, what you're really doing is generating a bunch of potential data sets, but if you run your analysis say five times, you will get five slightly different answers. They'll be similar but slightly different. With the multiple imputation strategy, you always get the same answer. So something that's easier to interpret.

The other difference between the two is multiple imputation, you can think about generating all these data sets. That can be very costly in terms of computational time versus maximum likelihood, you are running one model. So it will converge it into a little more computationally efficient and potentially easier to interpret.

There's always an interesting debate back and forth on a website missingdata.org. A couple statisticians blog back and forth about the best way to deal with missing data.

Tracy: Another question for the Alabama team: *You spoke of the case management claim. Can you talk more about what those are?*

CM: Those are maternity contractors that also in our current program pay for the ob-gyn delivery and prenatal costs, that kind of bundled care, plus a couple lab tests, plus case management and that kind of bundled service. That comes in as a specific claim for the delivery, and there are some rules around how they code that with a single data service for the date of delivery, for example. Then they also provide through a separate system a whole set of information based on somewhat like a "chart review." They'll go through the physician's record and find a lot of the information about the pregnancy and bring in the notes from their case management. That's the claim we get from them. It's a specific set of contractors in the 14 districts.

DN: It's not a true encounter claim. It's a physician-based claim almost but just from the maternity contractors for a global fee.

Tracy: Related to that, when you were working with case management data, can you talk about some of the specific gaps you came across? Cost? Utilization?

CM: The biggest gap we have with our case management program is it doesn't cover all of the deliveries. There are two major exclusions in it. One is mothers who are high-risk who are being cared for at our teaching hospitals across the state. Those are excluded from this case management system. So all the data

we get, and I'm a data guy, so all the great data we get on them we only have the claims data and not all this supplemental data coming in for those mothers who are considered high-risk.

The other piece is mothers who may not be fully eligible or don't receive any prenatal care. This could be mothers covered by third-party insurance or who, for example, are noncitizens and receive only emergency services at time of delivery and are not eligible for prenatal care. That works out to be a sizable chunk of our data and comes off of those. That's the biggest gap in the case management piece. We're able to get all the cost information from our actual claims data so where we can match the mother and identify the delivery we feel pretty confident in that information coming in.

Tracy: From the audience for Tom: *How do you deal with bundled codes? My analyses have been complicated by not knowing the exact dates and number of visits.*

TF: This is a really sticky issue. I don't have a perfect solution to this. I would make some suggestions for potential strategies. If I'm understanding the phenomenon correctly, the providers are just billing for a bundle of services and they're not having to necessarily accurately track an encounter of this date and an encounter of this date because they're being paid a lump sum.

I would look for an alternative data source, for instance using birth registry data where you have more fee for service type structure to develop a predictive model as far as the number of visits and the timing of those visits. Then using that model to try to predict within your bundled data what a likely or probable number of visits are and when those visits occur. That would be one potential strategy.

The other strategy if you have access to some sort of electronic medical record data or provider data, that also could help inform that. I don't have a perfect answer because it is a difficult thing but I would look for another data source to develop basically a prediction model and then try to predict the cases from there.

CM: I'd add something to that. We are in the process of modifying some of our claims procedures to require physicians to include in different parts of the claim the date of the first prenatal visit on a bundled payment code and the number of prenatal visits in certain parts on the claim form they submit so I can pull that off. I don't have enough data to know how well they're doing on it but we had sent out alerts and worked with our MMIS provider to start requiring that. As a payer we have that ability that other researchers do not, but that is one solution.

Tracy: This could go to all speakers, too: *Is Medicaid using ICD10 diagnosis and description codes yet, and if not, is this a reason for missing data? Tom first?*

TF: This is a good case of data missing data not at random. There are a couple different factors in play. One, providers haven't been totally accurate and comfortable with coding for ICD9, submitting claims for ICD9. They're not quite as familiar with ICD10, which is a very different structure. Also many of the contracts, especially within Medicaid, haven't been fully modified or vetted to make full use of all the ICD10 codes so they're more using the basic set of codes. That would be really what's happening.

So you're seeing this transition between the two systems (ICD9 and 10) and the underlying financial structure hasn't quite caught up to the detail you can get with ICD10 coding. I'd be really interested to see what Alabama has been experiencing with this as well.

DN: We keep a file from 2010 through last year so we are using both ICD9 and ICD10. One of the things we found as went to ICD10 was actually that's where a lot of errors were occurring. The example I gave

where we found a mother who had had a claim that hit the rules showing it was a delivery where it was morning sickness but they coded and used morning sickness with delivery, is one of the ones coded in the quality measures as a delivery. But as we looked at what was really happening with mom, that wasn't really a delivery. We think that was just a coding error probably associated with ICD10 and somebody not reading the full description. We've had other cases like that. We've had those issues coming in.

We do utilize some of the ICD10 information for birthweights and so if we don't get the exact number we would really, really look we can kind of get some ranges and gestational age I believe from there. So we take advantage of some of those pieces, but there's also some additional confusion that comes in from those.

Tracy: A question for Alabama: *How long did it take you to complete your maternity analyses from start to finish and what kind of staff skills did you have to bring in to get it done?*

CM: We kind of did a rewrite of this whole process early last year as we were finding some of these issues. We have a really good team. We have technical folks that are database administrators. We have some analysts and some clinicians on staff there were able to huddle up. It wasn't a full-time process for them but it was a significant project that took 4-5 months from start to finish, but the resulting work now is something we update on a monthly basis where we add the most recent data that comes in and then on an annual basis kind of merge that with vital stats. We've got it built now where we can update probably an 8-10 page maternity report that you saw several examples from within less than a day. So the data just stays up to date. In fact, this morning preparing for this I ran some queries off the table we have just already run, and we put the energy and effort into building the underlying infrastructure so we can do the analytics quickly and easily now.

Tracy: One more question: *Tom, you mentioned several SAS procedures and R packages. Anything for Stata users related to missing data?*

TF: Yeah. I will admit I'm pretty ignorant about Stata in general but I know there are a lot of health commands within Stata. The first one is MDESC. That is actually a summary function to summarize missing data. We talk about that monotone type pattern, it can help you identify that pretty quick. There's the multiple imputation strategy within Stata, and within their ML (maximum likelihood) there's also some options implementing maximum likelihood for missing data as well. And I believe there's a really good source for this, too, out of UCLA. If you go to the Stata permanent at UCLA they have some really good online resources for pretty much all kinds of data. Hopefully that can get you started.

Tracy: Thanks. Back to Jessie for key takeaways.

JP: To summarize some takeaways from today's webinar:

- We demonstrated how ignoring (deleting) missing data may lead to incorrect conclusions.
- We walked through the concept that an appropriate strategy to address missing data is determined by its *pattern* and *structure*.
- We demonstrated via Alabama Medicaid one method for minimizing the presence of missing data.
- Ideally, one would address completion at the data point of collection, but as this is not always possible, our next webinar in September will highlight analytical approaches to missing data problems.

Thank everyone for participating. I would ask all participants to complete the post-webinar survey as your feedback is extremely helpful. Slides and a recording of this session will be posted on our IAP Data Analytics website and we will email all participants with the link. For more info on our IAP program or addressed to the Data Analytics team, you can reach us at: medicaidiap@cms.hhs.gov.