Medicaid Innovation Accelerator Program Data Analytics National Webinar

Data Merging and Integration Medicaid

June 6, 2019

Keith Branham (Moderator), Research Analyst on Medicaid IAP Data Analytics Team, Data and Systems Group, CMCS

Introduction

Keith Branham: Here is today's agenda:

- Introduction: I'll introduce our four speakers with three presentations.
- Overview of the Medicaid Innovation Accelerator Program.
- Review of Data Merging Processes.
- Washington State's Experience with the Social and Health Services Integrated Client Databases.
- West Virginia's Data Integration Experience Merging Mortality Records with Medicaid Data.

Our speakers are:

- Jon Busch, Senior Director, Government Health and Human Services, IBM Watson Health
- David Mancuso, Director, Research and Data Analysis Division, Washington State Department of Social and Health Services
- Tanya Cyrus, Chief Quality and Integrity Officer, West Virginia Bureau for Medical Services, and Suzanne Lopez, Director, Compliance and Reporting Privacy Officer, West Virginia Office of Management Information Services

A brief overview of the Medicaid IAP and how this webinar fits in. This webinar is produced and funded by the CMS Medicaid IAP. IAP is a cross center collaboration at CMS that supports state payment and delivery system reform. There are four functional areas in IAP and this webinar falls into the data analytics functional area, which aims to increase state Medicaid agencies' data analytic capacity by providing resources, technical assistance and other informational tools.

The goals for today's webinar are that states will learn about:

- Reasons to link data
- Common methods for linking data
- Common methods for working with "noise" in data
- Overview of probabilistic methods for data linkage
- Concerns and limitations
- Examples of the data linkage experiences in WA State and West Virginia Medicaid agencies

Our first presenter is Jon Busch from IBM Watson Health.

Discussion

Jon Busch: The basic question, why would anybody want to do this? Linking data is hard or at least it can be. It can also be relatively easy. Sometimes you need to do this because the research and policy questions you have are best answered by using existing secondary data in combination with the data that you have. This is usually easier than going out and say doing a pilot study or an original data collection.

The bit of downside of this is that the secondary data is usually collected for other purposes, such as administrative data on the hospital side, which many groups work with that kind of data. It's great data but it may not have exactly the thing you're looking for. It may not have exactly the identifier you're trying to link on. There are also similar data elements, not exactly the same, that are present in different data sets, so when you combine those two you can get more information out of the result. These can be used in combination to create data analytic sets.

Here's a hypothetical example. You have a vital records data set and eligibility data set. This is not HIPAA because it's me and not actual data. In the eligibility data there's a record that looks like it might be similar. I do get a lot of mail through John Bush but I'm not sure we're the same person. Date of birth looks similar but not exactly the same and gender's a match, but you figure you have a 50/50 shot at this one. So not totally clear you're dealing with the same person here.

So how can you do this? There are three basic methods you can use to link different data sets: Match merging, deterministic linking, and probabilistic linkage. These are essentially progressively harder. In match merging you're conjoining or you can think about this as a SQL join on a common key. You can do this in any SAS data or any common analytic package allows for this. It's simple, really efficient, but not always possible. In the example we had there's no common key.

Deterministic linking is a little bit more advanced because you're not linking on a single field but you can think about it in most examples as assigning points for matches on different things. Now we only have three variables on our sample data, and you can think about Social Security number if you find a match on that, or eligibility ID, something like that, that might be worth a lot of points. For matches on first name, last name, you basically just add up the points and then assign the match or make a match between those that have the highest point match.

Then probabilistic linking really looks at all the data, not just points, but it looks at the likelihood that any two records are the same. So it combines all available information so not just what's available but also the frequency of what's available, and we'll talk about this a little bit more. It's more complicated to set up but you can also identify links that may not be possible with other methods.

Back to our example. What might we do? Match merge is not possible here because there's no common key, at least nothing that matches except for gender. That's not something you'd want to merge on. Deterministic, you can't really match on name. The only matches that you have are year of birth, which doesn't really discriminate between people very well, and gender, which doesn't discriminate among people very well. So what could you do?

In this example I worked through deterministic linking. In this case you're merging each record so the record in vital records for Jon Busch with three different candidate records on the eligibility data. So the first record we mentioned and two other hypothetical records, which have different combinations of

name, date of birth and gender. I've just assigned some arbitrary values to this illustration, but you can think about these as being generally something you may want to use.

In the first case we have a match on first name, sort of. We have no match on first name in the second record and then we have a little bit of a match on first name. There's no match on last name in any of the records. There are different points for the date of birth match. They're off by different combinations of months and years. Then you see in the final field you get one point for gender match, that's arbitrary, but then gender nonmatch is worth negative 10 points. So basically the process in deterministic linking is a fairly elaborate one. You go through and find all these point matches, and in this case the most likely record based on the point system we've established is Jon Busch on the vital records linking with Jim Bess, because the first name is similar, last name is similar, eligibility date of birth is similar to vital records date of birth and gender is the same.

Probabilistic linking is more complicated, which is good news and bad news. The good news is that there are algorithms and packages available to do the work for you. Also, if you're the kind of person who looks at the equations and the math to do all this work, it can be a little daunting, but they're essentially fairly straightforward ways to at least think about how you do all this. It uses multiple criteria and establishes scores. The main difference between probabilistic linking and deterministic linking is how the points are calculated or the thresholds are set. There are agreement points determined by the data rather than by us in advance. Then weights are based on the discriminating power of each comparison variable expressed as a weight.

In our example before we didn't have a Social Security number. If you think about two variables, so if you have two records, they match on gender or they match on Social Security, it's going to tell you more if they match on Social Security number, simply because there are more possible values. So a match on gender is essentially a 50/50 shot, maybe a little bit more, maybe a little bit less. But a match on Social Security number just statistically is unlikely unless it's the same person.

Scaling is a little bit different than weight because it gives you a mechanism for modifying the weight based on the relative frequency. So you might have a weight you have assigned for a match on a last name, and that's fine. Maybe it's meaningful. Maybe it's a high value for a match on last name, maybe it's a low value. Then within that you can scale name so that if you find a match on Smith, it doesn't assign very many points to that possible match. But if you find a match on Wobbe, that's much more likely to be the same person because it's not a very common name. So in the same way that Social Security matches are much more likely to represent the same person, a match on two records that match on Smith versus two records that match on Wobbe, Wobbe is more likely to be on the same person if there is a match made there.

The process for doing this, and we did see this a little bit before, you essentially join all the records. So you try and compare all the records. So you compare every record in vital records data with every record in the eligibility data. There would be a lot of records that don't link and that's okay. There would be some in the link group, and you can think about this a being a kind of distribution. So there will be lots and lots of records that don't link, that's fine. There will be a few records in the link group, and then there is a little bit of overlap in the middle. If you want, you can either set an arbitrary threshold and then put everybody above or below that threshold in the non-linked group or the linked group. You can also do manual review of the uncertain matches, that's fine. But most pairs are not likely to be linked.

Here we have an example. It's the same vital records data, same eligibility data, and then there appears a probability of a match. So this is the kind of output you might see from a probabilistic linking output. So it's not really saying that Jon Busch is John Bush, it's just saying that Jon Busch is more likely to be John Bush and Jim Bess is more likely to be a match than Jane Box. So there is still some threshold work to be done in terms of determining where the thresholds are for a match versus a non-match, but it can also be a little bit more straightforward.

The math behind this is relatively complex but you can break it down into some more intuitive steps and processes.

(next slide) One thing I really want people to take away from this is you don't have to start from scratch with a text editor and a stat package. Years ago when I first became involved in this we started to go down this road ourselves, and there's a lot of available research out there, a lot of published data. Now there are a lot of available packages so it can be used by anybody essentially to do this work, to do most of the heavy lifting for you. So you don't have to start from scratch with a blank SQL code or SAS code or what you have.

It's important to know your data. In the short example we used, what you have to think about, if you're doing a comparison between zip codes, it is more likely somebody would move across town and be the same person or would they move from one region to another? So knowing your data is important in terms of figuring out what the likely matches are.

There will be false positives and false negatives. One of the first times I was involved in something that would use probabilistic linking, one of the key investigators I was working with said it was very important to set the threshold match at 100% so that you'd only identify the true linkages. That's kind of missing the purpose behind this. It's almost unavoidable that there will be some false positives and some false negatives. It's not recommended to use for clinical decision-making, so if you're trying to match people with their medication history, probabilistic matching is probably not the best way to go. But this is a good way to look at patterns and trends in groups and the population as a whole.

A couple techniques that may be helpful in terms of working with imperfect data. Names often don't match, and we've been through some examples with that. One of the oldest algorithms available is called Soundex, that's a way to do it. There are other phonetic algorithms you can use. You can also use these not just on names but also on places. Then here's a quick step in the process, so keep the first letter and then drop all other appearances of vowels and then a couple other common letters and you replace consonants with digits. That's how it would work for Soundex.

You might think Soundex will solve all your problems. If you look in this case, Jon and John, they match on Soundex, Bush and Busch match on Soundex. But you've kind of got the other problem. You've got lots of other combinations of names that still match on Soundex but seem to represent different people. So Jim, Jane, Jon, and John, different spellings, all have the same Soundex. Bush, Busch, Box and Bess all have the same Soundex. So just knowing that somebody's a J500, B200 doesn't really tell you a lot about who that person really is. So know the limitations of the algorithms you're using.

So Soundex is good for some things, but there's an algorithm known as Metaphone on NYSIIS, which was originally developed for New York. You can also use what are called string comparators. This has the Metaphone value for the different names. So even though they're all J500's, you have JM, JN and a few

other JN's. But then on the last name Metaphone does discriminate a little bit more. You have BS, BSK, BKS, and BS. So all the first names, or most of them, still have the same Metaphone but very different for the Metaphones with the last name.

String comparators, this goes back to some work originally done for the census quite a while ago. You may come across those names. But what you're doing here is comparing any two strings—names, cities, what have you, and the basic method is to count the number of insertions, deletions or revisions required to make two strings identical. So Jon and John, you can create one from another, either by adding a letter to Jon or deleting a letter from John. So you can think about that as they're one letter off. John and James, however, that's a little bit different because you can either delete two letters or insert four. So it requires more work to get from John to James than from Jon to John. So in terms of a string comparator Jon and John represent a higher degree of similarity than Jon and James.

Linking records may be relatively simple, in the case of match merging, or somewhat complex with different degrees of apparent accuracy. It's an open question whether or not the records match that we have here are really the same person. We'll assume they do seem to be. There are a lot of existing methods and existing tools available. We developed some a long time ago and they have been used and improved on.

There's no single best method available. If you have five minutes and need to merge some data then maybe match merging is the best way to go. If you're trying to do a more extensive research project that has a timeline of months or years, then it might be worthwhile to invest in more advanced probabilistic matching. The probabilistic approach can seem a little bit complicated, a little daunting from a math statistics perspective, but you really can divide this into a bunch of things, a bunch of smaller steps that are really somewhat intuitive. So is it more likely that two records are going to match on Social Security number of based on gender? Is it more likely that records will match based on date of birth or zip code? And it may vary within your population. If you're dealing with schoolchildren like a pre-K population, they'll be restricted in range in terms of when the dates of birth are, so year of birth may not be all that relevant if it's a one, two, three or four-year range. So you need to know a little bit about working through data.

Same with comparing strings. There's no single best method to compare strings but you can take a Soundex. There are a few things out there you can rely on. The algorithms themselves are not terribly complicated and it's usually pretty easy to understand that you're going to take a version of this name and just condense it and distill it down into phonetic representation.

David Mancuso (Washington): We do operate a fairly large-scale integrated data system that does link data from dozens of data systems that are generally administered by our department or by partner agencies within the Washington State environment.

(slide 24) My presentation will focus on why we integrate data, so the business case for why we do what we do. I'm going to talk a little about the legal framework that supports how we do this work in a multiagency environment. I've got a few highlighted use cases and a link to our main research resource site that I encourage folks to take a look at if they want a sense of what we do more broadly. Then I'm going to close with lessons learned that will include a little bit of detail about our linking processes. John's done a great job giving an overview and our story would be very similar in terms of key aspects of the technology, but I'll give a high-level overview and point folks to a free SAS-linking software package

developed by somebody who currently works in our division, Kevin Campbell. I've got more content than time so will fast forward in several places. Let's move to slide 26.

So the business case for integrating data across social service and health service delivery systems and other data systems, employment data, education data systems, is that there is huge overlap in who is served by or involved in these delivery systems, and there are profound implications that flow from the services provided in one delivery system or potentially needed but not provided in that system because of underfunding or lack of access or challenges with engagement. Huge implications that flow from services that one system is accountable for and costs incurred or outcomes experienced in other delivery systems, and those cost and outcome implications are really things that speak to the impact that these systems have on people's lives and their quality of life.

This slide we're looking at has a lot of stuff on it, but it's really intended to illustrate the overlap between the three major agencies. There are other agencies, like the Department of Health and Department of Commerce, as well others I should be referencing, but the three agencies that deliver most of what we would consider to be the universe of social and health services that stage agencies deliver—Medicaid, cash and food assistance, child welfare services, in our state these are now delivered through three different agencies. The integrated data system we steward here in my division was something that grew up at a time where these delivery systems essentially existed all within a single agency, which had had some advantages in the development of this capacity. But it is a capacity that we the state are maintaining and building on as we've moved to this multi-agency environment, which may be more similar to the environments present in other states. This gives you a sense of the overlap and speaks to the importance of what can be learned by integrating data across the IT systems that support the different service delivery systems that these agencies operate.

On to slide 29. I want to get into the legal framework that underlies how we do what we do. Linkage, the mechanics of linking, is a really critical part of our master data management, that identity management component to it, but foundational to all of this capability is the legal authority and contract structures that allow us to access data that other agencies own or our partners within our department own that give us the legal authority to do what we do and define what we can and can't do with data that we're integrating. To understand the legal framework it's important to understand the nature of the work we do, and that although my division is called the Research and Data Analysis Division almost everything we do is really a business operations function as opposed to research. We're not doing most of our work to create generalizable knowledge, we're supporting business operations for the programs of our department and under contract with our partner agencies across the state.

Slide 31. If we want to frame what's the foundational legal challenge to do what we're doing it's how do we establish the legal authority to access identified, specially protected data and substance use disorder data provided by 42 C.F.R. part 2, child abuse and neglect data protected by CAPTA are a couple of many examples of this specially protected data, so this is not just HIPAA-protected data?

How do we establish the legal authority to do that in the context where we're performing business operations functions in a multi-agency environment? The long story short is perhaps somewhat ironically we leverage research disclosure authority to gain access to that data in an identified form. We do that through an IRB-approved study that allows import of sensitive clinic data under research disclosure authority to do the identity management necessary to link information. We have a set of companion data sharing agreements and service-level agreements with agencies and programs who support specific

analytic functions that we perform, and those data sharing agreements define what we can and can't do with their data, and they are incorporated by reference in the context of our IRB-approved study.

In terms of the use of information by analysts, we're generally leveraging limited data set concepts to operationalized minimum necessary criteria as analysts do their work. Now in many contexts analysts may still have the authority to access identified data in the context of their work, but that's not generally the case, so the notion of working with data sets that have been stripped of direct identifiers but still have the granularity needed to do the work is central to our data management processes.

Slide 33. I skipped a slide that has some legal references that give you a sense of the definition of what a limited data set is. It's important to note limited data sets are limited in the sense that there are certain direct identifiers that can't be there and in many contexts analysts can't re-identify that data directly themselves, but limited data sets are not limited in many other key ways. They can contain highly granular geographic information, highly granular time information. These are not the same as the Safe Harbor de-identification standard in HIPAA. These are still data sets that would be considered protected health information. Because they can be so granular, they can meet most of the analytic requirements of the work that our division does.

Slide 34. A couple things to emphasize here. One is that companion data sharing agreements and service level agreements with data owners are central to the authority to do the work that we do and are part of a larger governance framework. They define our obligations to data owners to do things like share information with them in a timely way to review findings before publication and a variety of other reporting obligations on what we're doing with their data.

Slide 36. This notion of data sharing agreements and doing work in partnership with data owners, that's really foundational to what we're doing, and it connects to authority to access data to integrate, to link, and then to analyze. Generally the work we're doing needs to be on behalf of the data owner or on joint behalf of the multiple data owners whose data we're analyzing. That connects to a notion of stewardship or partnership in the way that we do our work.

Slide 38. There's a link here if folks want to see more broadly the range of analytic work that we do. That link will point to hundreds of policy briefs and research reports that most commonly use integrated data from multiple sources. I'm going to highlight a couple things very briefly and move on to slide 39. So these are analytic products that are using integrated data. In the case that we're working with here it's MMIS data linked to carve out behavioral health data to profile outpatient ED and inpatient admission utilization on a per thousand member month basis applying a lens of behavioral health risk.

So you've got three columns here, looking at the overall adult Medicaid population, the subset with serious mental illness, and the subset with co-occurring mental illness and substance use disorder. These are partially overlapping subgroups. So this is an illustration of using integrated data to get a more comprehensive view of the relative risk factors in the risk profiling that creates these risk groups here, using the healthcare data to measure what's being paid for out of the medical assistant's side of the house, and showing that there is a pretty extreme gradient of risk or utilization as you move from looking at the overall Medicaid population to the subset with co-occurring mental illness and substance use disorder.

Next slide. This uses the same behavioral health risk lens in the same population but looks at outcomes coming from other linked sources. We're linking to information about housing instability that comes from

the cash and food assistance eligibility processing system, which actually does get passed through into our MMIS system, so it's kind of cheating for me to refer to it as linkage. But the linkage to statewide arrest data and employment data from the UI wage system, those are real external linkage processes that come through our identity measure process. One thing I would highlight here especially is you're looking at the directionality of what's good and what's bad—homelessness and arrest risks look very much like ED utilization and inpatient risks. So as you link these kinds of information you start to get a fuller sense of the impact of behavioral health risk factors in driving, not just "medical service utilization" but also these other adverse social outcomes of housing instability, homelessness, criminal justice involvement, which also present barriers to employment.

Next slide. The second and last illustration of use of linked data I want to give here is operationalization of average childhood experience concepts using linked administrative data. These are where we're using linked child welfare data to identify involvement in the child welfare system, linking to mortality data, criminal justice data, information from child welfare and cash assistance programs that give us a perspective on domestic violence risk. And again linkage of behavioral health risk information and health service information were generally from MMIS.

We're also using linked birth certificate records to create family relationships and child support enforcement records to supplement information for children not born in Washington State.

Next slide. We pull this information together and the story this slide is telling is which of these adverse childhood experiences are most strongly related with the likelihood that the child on Medicaid develops a substance use disorder by the time they're an adolescent. We can see that children which have experienced abuse and neglect have roughly speaking four times the risk of developing a substance use disorder by the end of adolescence relative to children enrolled in Medicaid who don't have that risk factor. So illustration again of the kind of information that can be derived from linked data.

To say a little about our lessons learned about linking, at the scale of linking that we're doing we do need multiple technologies, in particular because we have sources with varying linkage quality, and we do have a master linkage process that much of our information flows through that is centralized across many data system, especially our criminal justice sources. We have more challenge in identifying quality, greater use of aliases, and so we maintain satellite linkages around that.

Among the things to think about or opportunities to consider, we do have some IT systems, notably our MMIS system, that have tremendously accurately linking that happens internal to that system. It gives us an opportunity to leverage systems' identity management as a golden record concept to help further define what comes out of our core linking methodology.

We also have a context where we link to birth certificate records to help further refine identification in the case of multiple births, which can be particularly challenging because you've got the same birthdate, the same last name, and if not the same gender you might have a very similar first name. There are areas where we leverage that and frankly we'll be doing more of that leveraging as we continue to improve our linking going forward. I do want to highlight that there is a free SAS software package developed by Kevin Campbell when he was with the Division of Behavioral Health and Recovery, which is part of our Health Care Authority. He is now part of our division and involved in our linking here. He has a free SAS software package. In fact, a link didn't get embedded here; feel free to ping me for that link. The list of features of that technology reflects a lot of components John talked about in his presentation. It's free. The SAS

license is not free and the status to do it is not free. But it's really a powerful tool if you have the ability to work with SAS in your environment.

Just a few additional thoughts about lessons learned. Linking is complex but in the larger scheme relatively straightforward in terms of a functionality. Far more challenging in terms of doing good analytic work with linked information where you are linking to new data sources is developing real analytic mastery for these new data sources. And if you're doing this in an environment where you're linking data or seeking to link data with an external partner, it's important to understand that their data is complex. If you think about the complexity of your data if you are in a state Medicaid program, the complexity of your MMIS data, a lot of that complexity is going to exist in a child welfare system or a system supporting the TANF program or a system supporting the assessment for eligibility for home and community long-term services and supports. So linking data, identity management, hugely important. Complex. Important to do it well.

Developing the mastery to use what you are now linking to is something to not underestimate what that requires in terms of investment and what might be required in terms of building credibility with that external data provider to build trust that that data will be well-used and there will be a governance process that will ensure that more good than harm is done in that integrated data environment.

One more set of things about keys to success. Time has been important. We started a long time ago. It's been many years from the time that large-scale integration efforts started to where we're doing fairly broad, frequent, quasi-experimental program evaluation, predictive modeling, clinical decision support out of this environment. For folks who are beginning this capacity and moving up the mountain of building this kind of capacity, having realistic expectations about the level of resources required, the timeline for capacity development, is really important. I will pause with that.

Tanya Cyrus: It's my pleasure to present with Suzanne Lopez West Virginia's Medicaid experience with the IAP in the integration of mortality data in our data warehouse.

(next slide) Our presentation will touch on why West Virginia Medicaid decided to integrate mortality data, our work group and aspects of the project such as integration process, challenges and lessons learned. We will wrap up with a little information on what we are doing now with the mortality data and any next steps.

Why mortality data? Our initial response would be because we needed the cause of death for our quality measurement. However, as this slide states, those who have passed speak to us through their mortality data so that we can improve the quality of healthcare for the living and future generations. One example from history that supports this statement is Florence Nightingale's analysis of mortality data during the Crimean War. Through the collection and review of mortality data, Nightingale learned that poor sanitary practices were killing more British soldiers than the battle. Her analysis led to healthcare practice improvement in that unsanitary and unhealthy environment, and that ultimately led to saving the lives of the soldiers.

This slide goes into more about the transition to the IAP project to integrate mortality data. In early January 2018, we were discussing different data sets for integration. We considered four at that time—mortality data, inpatient hospital data from the West Virginia Healthcare Authority, claims data from the West Virginia Public Employees' Insurance Agency and data for incarcerated individuals who may become eligible for Medicaid upon parole and release. After many discussions about the different data sets, in

mid-January we decided we would submit an expression of interest form to CMS for this technical assistance for data integration with all four data sets listed because we could not pick one at that time. Following a few emails and conference calls with the CMS Task Six Lead, we had a conference call with CMS Task Six representative, we were notified of our selection for data integration technical assistance in late February. At that time the data sets were narrowed to mortality data, and in March 2018 we began the weekly conference calls for the project. Our target date for completion was September 30, 2018, which gave us six months to complete everything.

The next step within the project was to identify the data needed for the integration. Obviously we needed the mortality data. So early on we included the staff from the West Virginia Health Statistics Center. We learned that the mortality data from death certificates was available back to January 2011. The data was in two different formats, however. What was termed as the old format for death certificates was used until mid-2017 and that data had 60 data fields. The new format which followed had 136 data fields, or a little more than double that of the old format. Other data available to us for the project was the West Virginia Medicaid eligibility data, our West Virginia Medicaid claims data and utilization management vendor data as well as information from our West Virginia Department of Human Health and Services sister agencies.

We basically used the eligibility data and the mortality data. This slide gives you an idea of all the workgroup members we had involved in the project after our data needs were identified. You heard that early (go back a slide) on we brought back the West Virginia Health Statistics Center staff. At the time of the project I was with the West Virginia Office of Management Information Services as the director of quality and analytics, and I was the project lead for this project. Suzanne Lopez, the Director of Compliance and Reporting, who you will hear from next, was co-lead. We included our staff as well as the Medicaid Director of Program Integrity and Director of Quality management as well as the onsite data warehouse staff and the CMS Task Six lead.

Once the workgroup was in place, the data fields in the old and new formats were reviewed by West Virginia Medicaid to determine which would be included in the data warehouse. Nearly all were selected, and the old format was used for the 2017 data, which was the one that was split under the two different death certificates. The overarching goal of the workgroup was to match the mortality data with Medicaid eligibility data with a match rate of at least 80%.

To detail the specifics of the data integration, I will turn to Suzanne Lopez. (next slide)

Suzanne Lopez: In the next few slides I'll be discussing our matching strategy, matching results, as well as some challenges we faced here at the project. When we started examining the mortality data, our first thought was to take a look at the existing strategies used for matching birth data and that match exceeded 85% for the child and 96% for the mother for that project. DHH Medicaid tried several strategies. However, the most effective was to use the mother's Social Security number to match Medicaid claims data to the birth certificate and then match Medicaid eligibility start and end dates to determine if the child was born during that particular time frame. However, it didn't take us too long to conclude that we couldn't go that route for that mortality data match.

So we then tested the five strategies you see on the slide, and of those, we chose number five, Social Security number, date of birth, and gender. We found that this combination brought us the highest match rate and lessened the potential to create any soft __[00:53:10].

(next slide) This chart is pretty self-explanatory. It identifies match rates from the initial load, which contained mortality data between 2011 through August 2018. As you can see, the match rate during the project time frame using Social Security number, date of birth, and gender gave us an overall match rate of 87.96%. With this initial load we have continued with monthly loads and the data warehouse currently holds mortality data up through May 2019 with the current rate that fluctuates between 87% and 89%. So we definitely exceeded our project's goal of an 80% match rate.

(next slide) We were faced with several challenges during this project. As Tanya mentioned, the death certificates were comprised of two formats, the old format and the new format, and it did take quite a while for the workgroup to choose the data field we wanted from the new death certificates. Once we did, though, IBM had to work with Vital Statistics to match those data elements from their system to the field in the data warehouse.

Another challenge we faced was relating to mortality data of West Virginia residents who die out of state. At first we planned to include out-of-state data but we quickly found that that is not readily available nor is it comprehensive. We would not have been able to receive any type of automatic data feed, and we also found that states will not readily release their death data to other states, and even if they do, they may not provide a complete record to the requester. West Virginia would have had to send a request for that death certificate data and ask that state for the specific information we needed for each person or record, and that includes names, date of death or cause of death for each personal record, and that could include names, date of death and/or cause of death for each person.

The next three challenges on the list are related. They all deal with both system and time constraints. As I mentioned, matching the data field took a while, but we also ran into some issues in setting up the monthly fee. We spent more than anticipated determining a feasible schedule for the ongoing data loads. A formal change request is required for any project that goes above and beyond the contract responsibilities. At that time, the plan was to include the initial data load and the monthly load. But since we only had six months to complete the project, we ended up having to do two CR's, one for the initial load and then a second one for the ongoing monthly load.

That change request process is normally pretty straightforward, although it can be lengthy. We spent additional time working with our purchasing office because we were trying to determine whether this would be a no-cost CR. Because of our first experience with a no-cost CR, and everyone that was involved in the project was a little unsure of what that process entailed. Although we finally concluded the CR would be no cost, our biggest hurdle was yet to come. At the time of the project, luckily or unluckily, our vendor was going through a name change, which created further delay with the purchasing process, and all CR's that contained the old vendor name were put on hold. We were very lucky, though, because this did not apply to the no-cost CR, and even though we could not load the data into production at that time, we were able to load it into the test environment prior to the project deadline. So even though we did face challenges throughout the project, we did count this as a success.

Tanya Cyrus: So you've heard about our challenges and with those come the lessons learned. Early lessons learned:

- Identifying only one data set for integration saves time at beginning of project. We spent a few weeks actually working with CMS and the Task Six group in order to narrow down our data set to the

mortality data for integration. It will just save time if you can narrow the options at the beginning. It will provide you with more time for a project.

- Mortality data itself may be incomplete or it may be delayed for weeks or months due to activities such as medical examiner determination, review of the death certificate for coding by the CDC, or a lengthy legal investigation.
- One of the most important lessons learned that really goes without saying, we benefited greatly by bring the data warehouse onsite staff into the project at the beginning. And the CMS Task Six Lead having remote access to the data warehouse really helped us with expediting communication about the data and the resolution of issues.
- Another lesson learned not on the slide is related to phonetic name matching. We did not have that kind of software available to us for the project but we did learn that in the future that is something we want to use for matching strategy.

(next slide) That takes us to next steps. As Suzanne mentioned we have accomplished the first activity listed, the monthly updates, and we are current in the data warehouse up to last month. We are in the early phases of analysis and determining next steps from what we are learning from the mortality data. Just recently we have provided data to our medical director to evaluate an external opportunity related to an infant mortality reduction project. We are also in preparation to propose a quality improvement project to our Office of Managed Care to collaborate with the Medicaid managed care organizations on reducing mortality within 30 days of a hospital admission. This last activity listed, pursuing mortality data for West Virginia who die out of state, is a bit more time-consuming, but it's a goal we're going to leave on the list of things to do.

That concludes our presentation on the West Virginia Medicaid IAP Project for Mortality Data Integration. This slide includes Suzanne's and my contact information. 304) 356-5402, tanya.c.cyrus@wv.gov; (304) 356-5170, Suzanne.p.lopz@wv.gov. In closing I would encourage other state Medicaid agencies to take advantage of the IAP process. It is a state forward application process and the workgroup as a whole feels that we could not have accomplished as much as we did within that time frame of six months without technical assistance. If you have not considered integrating your state's mortality data into the data warehouse, I encourage you to do so. We believe that we will learn much from how West Virginians die, and that information matters in terms of population health.

Keith Branham: Now I'll turn it over to IBM to do the Q&A session.

Q&A Session

Padmaja: A question for Jon: *Why doesn't everyone use probabilistic linkage?*

Jon Busch: A few reasons. It's a little bit more complicated to use. I give the example if you have five minutes and a common identifier then match merging is fine. If you have an identifier that you trust, for example, so if you have a state ID or eligibility ID or something like that, it's faster to merge on that. It's a little bit more straightforward. And there's no real uncertainty in that.

If you introduce the probabilistic process then it generates a whole bunch of nonmatches, some matches and then potentially some uncertain matches. I'm glad that linking was mentioned because I'm personally a big fan of that. I think that's a great use of some of the commonly available algorithms out there, as I

think you mentioned, and I don't get any money from it by the way. But you do need SAS and a person to do it, but otherwise it takes care of a lot of things for you.

Padmaja: *Can you give an example of sources of vital records?*

Jon Busch: Presumably these would be vital records within your own state. I use vital records as an example. The mortality data would be one actually. So any kind of vital records data within your state organization. As I think we covered here a little bit today, these data are collected for different reasons. Often when the data collection process is set up, not everyone is consulted in terms of who might end up using this data downstream. So this is why it's kind of helpful to have methods for working with kind of noisy data and imperfect matches in the data.

Padmaja: *When is it okay to use simple methods rather than probabilistic or deterministic matching for linkage?*

Jon Busch: Anytime you're okay with it really is kind of the right answer. All of these involve relative costs in terms of labor, so again if you have a relatively straightforward match or straightforward identifier that's common across a few data sets, there is probably not much of a case to be made for doing anything more complicated than you have to. Especially if you're doing it on a regular basis. If you're doing an annual study or something like that, in terms of an epidemiological research or something like that, it might be worth investing the effort to try and find matches you might not otherwise be able to find, either because you don't have a common identifier.

Even on the match rate for the West Virginia example, the rate they were aiming for was over 80%, which I think is really good. So if you're happy with a match rate of 80% and they actually got a little bit closer to 90, so if you're able to exceed your goal match rate then you've done a good job with whatever you're trying to do.

Padmaja: David: *What are the high opportunity areas for data integration and linkage across social and health delivery systems?*

David Mancuso: I almost hate to answer this because I don't want to give short shrift to other areas, but I do think there are a few areas that really stand out in terms of the scale of implications for impact on lives of individual served or folks who need services, and the scale of impact. This would include for those states who may have carve-outs and challenges from the Medicaid program perspective accessing or from the medical assistance program perspective accessing behavioral health data, integrating mental health and substance use disorder service data with the medical side of the house, that would be in the highest sphere of potential value. Linkages of behavioral health data with criminal justice data. Linkages of child welfare data into an integrated Medicaid physical and behavioral health data environment.

Those are some of the ones I would put in the top tier of potential impact given the scale of the interactions among the delivery systems, the proportion of kids on Medicaid who have some interaction with the child welfare system over the course of the year, for example, and the implications of those interactions for the healthcare cost trajectories, including the behavioral healthcare cost trajectories for kids on Medicaid, really profound impacts, so those are some of the ones I would highlight.

Padmaja: David: *Are the data sharing agreement templates you reference available at the access online?*

David Mancuso: Not online but I would be happy to share templates. Through the folks convening this webinar I'd be happy to follow up and share examples.

Padmaja: *For someone looking to start to build data linkage capacity, how large of a team and what staff skills would be needed?*

David Mancuso: It's a little bit tricky because of thinking about what the scale of that linkage activity might be. If I were to think of it in terms of potentially multiple systems potentially involving not just data internal to your agency that might be siloed within the agency but needing to negotiate and build relationships with external partners, I think you could do a lot with something like a 5-person team and really build, in terms of building a foundational capacity, that could in time, with time to acquire access to data, to do linkage, to do that with quality and things that relate to concepts like validation, build a relationship with external partners, with a 5-person team one could do a great deal.

It still would be a small team. That size would have the capacity to look into a small set of the challenging issues that can be explored or interesting areas or impactful areas that can be explored more generally across social health service areas, so having realistic expectations about what a team even of five could do would be important, but a lot could be done with a team that size. I do think you need who leads that team. We tend to focus on quantitative social scientists as the skill set, the training to lead this work. Often they're trained in the software that's relevant for analysis like SAS or other software. They often bring a real passion, a real intellectual interest in the subject matter, which is going to be really important because you're going to need folks who really want to dive in to understand what messy, complicated data means in a complex policy and program environment.

Padmaja: Tanya and Suzanne: *What other data sets did West Virginia Medicaid consider for integration and what data set would you like to work on next?*

Tanya Cyrus: The four data sets we considered were: the mortality data; the West Virginia public employees' insurance agency claims data, which was really linked more to the all-payer claims database that we're working on; the inpatient hospital data from the West Virginia Healthcare Authority; and the data for the incarcerated individuals who might have become eligible for Medicaid at the time of their parole or their release.

As far as next data set for integration, we really haven't pinpointed one but I know we're doing a lot of work here with some of our sister public health agencies. I would guess we may integrate data from the Office of Direct Control Policy related to the substance use disorder epidemic in West Virginia.

Padmaja: Tanya and Suzanne: *Which entity or department conducted the linkage and what software did they use, and were records without Social Security numbers excluded from them?*

Suzanne: The entity we partnered with, our vendor was IBM Watson. We also worked with biostatistics, which is within the Health Statistics Center within the Department of Health and Human Resources, the office I worked for, the Office of Management Information Services, and then West Virginia Medicaid.

Tanya: The technical assistance provided to us through the IAP project, that was the piece for the linkage, and that's the piece we didn't have available to us.

The records without a Social Security number, they were excluded. Most of those were for the babies that died at birth. What happened is because of the monthly updates once a Social Security number is available that information is added to the mortality data that's loaded into the data warehouse.

Padmaja: David: *How do you de-identify the data you use from all the different systems in the IDS?*

David Mancuso: We don't de-identify it and that's kind of a term of art. I'm being maybe too clever by half with that. So the limited data set is not de-identified. It's stripped of direct identifiers. It's still granular enough that it would be considered PHI under HIPAA. So I'm being a little bit lawyerly around that term. So there's an arbitrary integer identifier, so the ID if you will, that's used, but the data is not strictly speaking de-identified. It is in the limited data set form and there's legal authority to analyze it at that level of granularity, but this is not about getting data to something like the HIPAA Safe Harbor de-identification standard. That would leave the data not nearly as useful as we would like it to be.

Conclusion

Keith Branham: (next slide) A few webinar takeaways:

- Medicaid beneficiaries are often served by multiple social and health service delivery systems, which create many high-value analytic opportunities using data linked across delivery systems.
- It's good practice to try multiple matching strategies and use existing algorithms and research that best suit your data and needs.
- For linking mortality data, try to integrate all available data at one time to expedite predictive and prescriptive analytics.

Thank you for participating. Please complete the post-webinar survey. For more information about IAP, contact us at our website or at MedicaidIAP@cms.hhs.gov.

[end]