National Quality Forum

Moderator: Scientific Methods Panel
October 18, 2018
2:00 p.m. ET

OPERATOR: This is Conference # 4791375.

Female: Good afternoon, and welcome to the Methods Panel Subgroup Number Three Follow-Up Measure Evaluation Call. We want to begin by apologizing for the technical difficulty at the start of the call and for the delay. We apologize and we thank you for bearing with us.

So, we will get right into it. We will begin with the roll call this afternoon. Do we have David Cella on the line?

David Cella: Yes.

Female: Thank you. Bijan Borah?

Bijan Borah: Yes.

Female: Great. Matt Austin?

J. Matt Austin: Here.

Female: Jeff Geppert?

Jeffrey Geppert: Here.

Female: Mike Stoto?

Michael Stoto: I'm here.

Female: And Lacy Fabian. OK. We will check in intermittently throughout the call to ensure that Lacy has joined. So, today, we'll be casting votes using the same link that you used last week during our main call. As a reminder to everyone, this is a public call. But, there will be no opportunity for public comment, and questions cannot be directed to the developers.

So, with that, I think we can go ahead and get started on the two measures slated for review today. As a reminder, that is 3483 Adult Immunization Status and 3484 Prenatal Immunization Status. I'll hand it over to Karen Johnson.

Karen Johnson: Thank you, Miranda. So, we only have two measures to get through today, so we will see how we go. We – first of all the first measure, Adult Immunization Status – this measure actually, according to your initial ratings, would have gone down on validity.

However, we want to discuss validity on today's call as well mainly because of some of the similarities between this measure and the second measure. And we want to make sure that – it's fine for you to have different ratings and votes on these. But, we want to make sure that those different ratings and votes are because there's actually differences in the measures themselves. So, we want to make sure that we have consistency in terms of applying your – the criteria, et cetera.

So, let's go ahead and start with 3484 Adult Immunization Status. And bear with me. I should have already had this open. It's page – sorry – it's page 17 of your discussion guide. And this is a new measure. And it's a little bit difficult – or it was for me. I'll speak for myself. I was a little bit difficult (leading to) understand exactly how this measure is constructed.

So, we are treating it as a composite measure. And within this measure, there is – there is really five individual performance measures that are being put up for endorsement. So, one is Influenza, the second is Td or Tdap, the third is Zoster, the fourth is Pneumococcal and, then, the fifth is the composite measure that brings together those four. OK.

And the denominators of each of these components and then, of course, the overall composite is it's not people. So, it's not patients. The denominator is actually the total number of recommended vaccines. So, in a way, it's patients. But, it's really vaccines.

And I would imagine that the reason that this was constructed in this way is because the vaccines have different – not every one of them applies to the same group of people. So, for example, Zoster is targeted toward people 50 and older whereas Influenza I think is – I'm forgetting if it's 18 and older or not, but it's a different set of patients. So, I think that might have been some of the confusion about the measure.

It's not our typical all-or-none composite. And it's also not our more typical composite that aggregates completely independent measures rolled up at a – for a particular entity and then combined. So, it's a little different than either of those two.

Michael Stoto: Karen, this is Mike Stoto. I think you were being too generous to them. I don't think that this specification makes any sense, to be honest. In particular, the measure number five, they say is – the numerator is the sum of the numerators one through four and the denominator is the sum of the (numerators) one through four. I just don't think that makes any sense.

Karen Johnson: Yes. It's – I think if you think about it in terms of everybody that is kind of a target for the Influenza – and maybe you're talking philosophically it may not make sense. But, you – do you understand the math at least that they are putting forward, Mike? Maybe that's the way I should ask you first. And, then …

Michael Stoto: I'm not even sure I understand the math. I mean I – it may make sense in some place. But, I don't think that the statement of it – if you ask me to – if you gave me the raw data and asked me to calculate this, I don't – I don't think I could do it.

Karen Johnson: OK. So, that's fair enough. Did anybody else have any difficulties on this? I'm pretty sure Mike wasn't the only one. I think in different parts of the

submission, they kind of flipped between talking about patients versus talking about recommended vaccines. So, it wasn't completely consistent across.

Let's kind of keep going. This is a health plan measure. So, their data come from lots of different sources, depending on what the various health plans use. It is not risk adjusted. And in terms of ratings for reliability, definitely split across the reviewers. So, therefore, consensus was not reached.

And for validity, once again, split across reviewers but really tilting towards the "Does not pass validity." But, we wanted to pull it and talk about it because of its kind of similarities with the next measure, 3484.

And, then, in terms of the composite construction – and, Mike, this is I think getting to some of your questions but not all – we have a three "High" and a two "Low" split. So, again, consensus not reached. So, let's just start with our discussion of reliability. So, definitely some concerns about the specifications and the clarity of those.

So, maybe not quite understanding how the components are weighted. As I mentioned before, the descriptions of the units of measurement and the numerator and the denominator versus how it is described maybe in the calculation algorithm and then the description may be not so consistent.

And then, also, this question about continuity of enrollment and just maybe not quite understanding what happens and how you get your data if people aren't continuously enrolled and how far back the enrollment have to go. So, those were some of the concerns.

For reliability testing, for each plan type, they provided a median reliability estimate for the four component rates and the composite. And the values were 1.0. So, this is, of course, perfect reliability, which I think was a little surprising to some folks. And our question is is this because it's a function of extremely large sample sizes which – I don't know if we actually had the full sample sizes provided to us. I don't recall what they were.

The testing was done really on three plans. So, this was based on their field testing. And each of the three plans had all three of the product lines. So,

basically, they are stratifying this measure by commercial Medicare and Medicaid. So, it's a pretty complicated measure.

Michael Stoto: So, this is – this is Mike again. I have another hypothesis. It's that they didn't do the calculations correctly. If you look at also at the – at the next table about validity, every measure – there is either 1.00, 0.00 or 0.50. And, you know, numbers never come out that – (this thing) so consistently. I really don't trust the calculations.

Karen Johnson: OK. That's fair. Why don't we – why don't I stop talking. The things that we really thought were questions were some of the specifications and just understanding how the measure is constructed in the first place and these – the reliability estimates of the 1.0, again, remembering that they are – they are health plans but they are – only three of them using their build testing. So, anybody who would like to kick in? Mike, if you want to continue or if others want to join in?

Jeffrey Geppert: Yes. This is Jeff. I just concur with Mike that the reliability numbers – I mean if you look at the Table One, I guess they do give the sample sizes for the three plans, which are in the millions.

Karen Johnson: Yes. They are there. Sorry about that. I'd forgotten where they were.

Jeffrey Geppert: So, it's – I don't even – I don't know what that means. I mean, what a – what a – what a metric for like one, you know, free plan with three million – I don't know. I don't know what that means. Are they comparing across the three plans? That's not – I mean to calculate reliability, kind of you have to have – you have to have a certain number of measured entities.

And three is not enough measured entities. So, the number just – and I think if you did the calculations correctly, you pretty much can't get like a one. That's almost – that's – I think it's almost impossible. So, I think the numbers are meaningless and probably not calculated correctly.

Michael Stoto: Yes. This is Mike again. I think that the meaning – the important numbers is the number of plans, not the number of participants in each plan. And three is just no way you can get such number so close to one.

Karen Johnson: So – this is Karen – just asking real quickly, they say, I believe, that they used the (Adams) beta-binomial methodology. Does this sound right? They are – each component is – did this patient get this? And, so, it does sound like the (dichotomous) nature would work. What would you need to see from them? Like their output of the beta-binomial? Or would that – I mean what would – what would help?

Jeffrey Geppert: Well, that – I mean you have estimate these (shape) parameters. So, they could provide the (shape) parameters, the alpha and the beta.

Karen Johnson: OK.

Jeffrey Geppert: But, again, you need – you need – it's just like – sort of like (central limit), right? I mean you need to have like a certain number of entities in order for the statistic, you know, to be meaningful and (to reuse 2Q). So, they need – they need more plans. They need to take the plans they have and break them up into some meaningful way.

Karen Johnson: Now, is that definitely the case – I mean just kind of looking at their spread of their – the results of the three plans, they have a huge denominator and a big spread. So, yes, they definitely only have three plans. But, anyway, we're – I guess – forgive me.

I should not be speculating on maybe what might be going on. So, anybody want to mention anything else? I mean we have the – are there other questions about the specs that we would like to have clarified or that really concerned you?

Bijan Borah: So, this is Bijan. I was one of the persons that kind of questioned about how they actually ensure that – that continuity in enrollment because without really knowing it, I think for some of the measures, these patients or these sort of (unknown members) can have one of those five vaccines before coming into the new plan.

I guess I – it's not clear as to how they will ensure that if they don't really specify continuous enrollment going back to (depending on what they know,

what do they think) (inaudible) how many years? It's just not clear as to how they would ensure that that's the case and that (piece) may or may not have actually had it – (had) a particular vaccine before coming to the new plan or plans that are included in this particular measure.

Karen Johnson: OK. Other questions or concerns? A few people actually voted either "High" or "Moderate." So, is there anything that – and I don't remember who voted what way. Was there something that kind of tilted you in that way even though maybe some of the specs weren't quite as clear?

J. Matt Austin: So, this is Matt. I have voted "High" partly because of the very high statistics that they had resulted in for the beta-binomial statistic. But, after hearing the concerns of others about the likely impracticality of everyone coming out as 1.0, I would probably vote differently this time.

Bijan Borah: And this is Bijan. I voted "Moderately." But, I actually would not go for "Low."

Karen Johnson: OK.

David Cella: This is Dave. It sounds like we have a consensus. And I'm just wondering if maybe the calculation – the reason it's 1.5 or zero is because they just went at the level of the three plans. And so, then, you would get one if all were – if all were consistently, you know, rank ordered and you'd get 0.5 if one wasn't and zero if not – if two weren't – something like that. The point earlier being there is just not enough – not enough entities being compared to have a reasonable spread of reliability.

Michael Stoto: This is Mike again. When we talked about developing this new procedure, we entertained the possibility that we could actually talk to some of the developers. I don't know (if it would resolve) some of these issues. And I have flagged this one for that, but I probably did it too late – it could be – for that to actually work.

But, I think – I think that what – for me to be comfortable, I would have to talk to them and make sure that I understood what they really did and how these things were really – were defined.

Miranda Kuwahara: OK. Yes. It may have come in too late for us to make it happen for you, Mike. That's kind of the problem. It had to happen pretty early on in your review. And …

Michael Stoto: Right.

Karen Johnson: Yes. OK. We are going to keep going, which is we do need to talk about the validity a little bit and the composite construction. But, if you have your – so, it sounds like if the measure does not pass, then you'd like to see specifications that are more clear in terms of how the measure is actually constructed, more information about how enrollment works, the continuity of enrollment – that sort of thing – and, at minimum, the (shape) parameters so that you understand what went into the reliability calculations and how the ones came out.

I will tell you that we are not unused to seeing very, very high numbers at the health plan. They almost always come out extremely high, at least in the past, for other measures. So – and I – again, I always assume it was because of the really high denominator numbers. So, with that, if you have your SurveyMonkey open, if you would go ahead and cast your vote for reliability.

And, then, in terms of validity, I don't really have to say too much here. They assessed the validity of the measure using construct validation. So, they correlated the five measures with themselves – so, the five components with the overall composite. And, then, for commercial plans, they had the data to be able to correlate the components and the composite result with two other immunization measures, one for children and one for adolescents.

They provided us the hypothesis that they were testing against and the results. And the – within measure correlations, components with the composite, et cetera all turned out to be one for almost all of the combinations. When you look at the comparison of the results just between the commercial plans against those two external measures, the correlations were pretty high, but the directions weren't always in the way that one would expect and the correlations, for the most part, were not statistically significant.

And, again, they were using the three plans there. They did describe their methodology for thinking about and creating the measure and getting it approved. We did noted that it doesn't appear that what they did exactly conforms to what we are looking for face validity. If the new measures that we have – (their staff) had conformed to what we were looking for, you could have, you know, just kind of relied on that and went from there.

There was also some uncertainly about the extent of missing data and then kind of another kind of overarching concern about patient choice and how that is or is not kind of considered in the measure. So, with that, does anybody want to talk about any of those points?

David Cella: Karen, this is Dave. I'm sorry to break in with this question, but I'm trying to figure out how to get into the voting again. Is there a link?

Karen Johnson: Yes. Miranda, can you remind us?

Miranda Kuwahara: Sure. There was an e-mail distributed last Thursday, the 11th. I can forward that e-mail to everyone so that it's at the top of your inbox. And that will include the link to the SurveyMonkey.

David Cella: OK. Thanks.

Karen Johnson: And let me ask did we ever get Lacy on the line? Lacy, are you there?

Female: I did get an e-mail from – like an automatic e-mail thing that she's got a conference (and she might not be able to attend today).

Karen Johnson: OK. She might not be able to attend. OK. So, back to validity. Do you guys want to (inaudible) either the correlations or the methodology or any of the threats to discuss those?

Michael Stoto: And this is Mike. I already mentioned the problem with Table Three, how everything is either 1.00 or 0.50. I think I should – zero is in there two, but they actually none of those. Again – so, I don't trust these numbers having been calculated correctly. The good thing is they have a note at the bottom of the table where they talk about significant (inaudible) 0.5 rather than 0.05.

Karen Johnson: All right.

Michael Stoto: So, in any other – in any other situation, I would just regard that as a typo. But, in this one, I really am concerned about the presentation of the results in a much more serious way.

Karen Johnson: OK.

Bijan Borah: And I think – again this is Bijan – that would potentially be what Dave mentioned. I think it's just that they have only three commercial plans. And, again, I think for this purpose, that might be (too small a number).

Karen Johnson: OK. All right. Anything else do you want to bring out on the validity? OK. Go ahead and cast your votes on validity. And, then, apologies – in your Items To Be Discussed section, I should have had a bullet point because we do need to talk just really briefly about the composite piece.

They evaluated the composite, the construction by doing a Cronbach's alpha analysis of the various components with the outcome. And they explored potentially trying out the kind of typical all-or-none construction and they did some sensitivity analysis to examine what would happen if you include versus exclude various components. And they have that information for you.

And I did want to just remind everybody – because we don't have too many composites, I was going to bring up – bear with me. When you're thinking about the composite performance measures, you want to think about whether the measures fit the quality construct and add value to the overall – sorry – or the components needs to fit the overall quality construct and add value to the overall composite while being as parsimonious as possible.

And you also want to feel comfortable that the aggregation and rating rules are consistent with the quality construct but being it simple as possible. So, that's what we want you to think about when you look at the sensitivity analysis that they looked at as well as their Cronbach's alpha analysis.

So, with that, does anybody have any concerns or questions or anything about the composite construction? That one – again, it was a "Consensus not reached." Three people were very happy and two people were not so.

J. Matt Austin: (Inaudible) one of the happy folks. The criteria you read off, Karen – I thought the measure addressed those. It was parsimonious. It felt like each of the individual measures added something in terms of (inaudible) to a composite. And I thought that sort of checked off the criteria that you had listed.

Karen Johnson: OK. Thanks, Matt.

Bijan Borah: This is Bijan. I agree. I think that's the reason why (I) (inaudible) commented with (in terms of consistency).

Karen Johnson: OK. How about the folks who voted "Low" on the composite construction? Any sharing of what concerns you have on that part?

Michael Stoto: This is Mike. I've been so critical about the parts of it. I have to say I do think that what they did with the composite (inaudible) (seems to be balanced).

Karen Johnson: OK. Thanks, Mike. OK. I don't want to push it too much if you don't have other things that you want to share. I'll give you just another minute. OK. I'll go ahead …

David Cella: (I was just going to say) I can't find the – I can't find the October 11 link and I didn't get resent.

Miranda Kuwahara: It was – the e-mail was just resent. So, it may be in the (Ether) right now. But, it's on your – it's way to your inbox.

David Cella: Thank you.

Karen Johnson: OK. For those of you who can – and, Dave, as soon as you get your e-mail, go ahead and cast your vote on the composite construction.

Miranda Kuwahara: And the composite question was just added. So, for those of you who have submitted your votes for reliability and validity, you can go ahead and submit your vote. Re-click on that link, click on this measure and then submit your composite vote. For Dave, it sounds like you have not had an opportunity any vote. So, you can submit all three votes at once.

David Cella: OK.

Karen Johnson: OK. Let's go ahead to the next measure. So, it's 3484. And this one is prenatal immunization status. So, this measure is quite a bit simpler. And it is kind of the usual all-or-none type composite with only two components that are combined via the all-or-none methodology.

So, again, there's three individual measures being put forward for endorsement. One is Influenza, the other is the Tdap and, then, the all-or-none composite itself. So, there is just the three. Again, it's a health plan level of analysis. And the denominator – I think it's probably simpler this time around because it's the same denominator really for all three of the measures. It's actually the number of deliveries.

So, in terms of reliability, this time reliability passes with a "High" or a "Moderate" rating. So, this is a little bit different than where you landed in the last measure. Again, the data – let me make sure I am looking at the right thing, 3484. I believe they had five plans this time. And this time, it's stratified by commercial and Medicaid. So, this doesn't apply to the Medicare population, obviously.

So, they had three plans for the commercial side and two plans for the Medicaid side. So, I think the – even though people rated "Moderate" – this would pass with a "High" or "Moderate" on this one. It seems to me like it's kind of the same issue as we had in the last time. So, I'm wondering if you guys would want to rethink your rating on reliability for this measure.

There were a few concerns about specs about not nearly to the extent of the last one. So, that might have been the difference. The specs question was I think Bijan's question about continuity of enrollment and exclusions. I think there was little inconsistency about the exclusions whether it was less than 20

weeks or 37 weeks. Probably a type in there somewhere. But, I think those were the two things that were pointed out.

Jeffrey Geppert: Karen, this is Jeff. I am – I am re-evaluating given the concerns that were raised on the previous measure. Just estimating reliability on an (N of three) I don't think is sufficient or an (N of two).

Karen Johnson: OK. All right. With that, I think – I think we would be more comfortable at NQF just to ask you to revote on reliability again. Again, the same methodology. Very few plans but lots of people in the plans and, this time around, probably not quite as many concerns about specifications. Anybody else like to add anything on reliability?

Male: No. (Not here).

Karen Johnson: OK. So, with that, if you would go ahead and reopen, right, your survey and go ahead and cast your vote for reliability. And, then, for validity, it was kind of the same idea here. They did some score-level validation of the measures.

Again, for the – they correlated the components with themselves and the overall composite and, then, for commercial plans only, correlating the results of the components and the composite with a childhood immunization measure as well as a measure of prenatal and postpartum care. So, again, two separate kind of external measures.

They described their hypothesis that they would expect. And like the last measure, the (measure) correlations were high but not statistically significant. The correlations between the various – the other two external measures moderate to high. There's a few exceptions there but, again, not always in the expected directions and, for the most part, not statistically significant.

Again, they described in their face validity assessment that it has a (conform to error) requirement there. Again, it seems like there were some uncertainly about the extent of missing data. And I think for both of these measures – and I – and I apologize I didn't mention it before – this is – these are new measures that they plan on doing more comprehensive analysis on these measures after they have been implemented for a year. So, it sounds like they

haven't been implemented. And that's probably why you received (field) testing results.

And then, finally, a few threats that were – the (just had) to do with the enrollment and properly accounting for the immunizations. And I think (it's) probably a little bit more concerned with the Medicaid plans and the (churning) possibility and then the question about patient choice. So, does anybody have anything to add on validity for this measure?

Male: At first, I had to sort of double check that this wasn't an e-measure because the sort of three sort of stuck out, you know, like if you were developing an e-measure, you'd have like three testing sites and this basically all you could do. But, it's not.

So, it's – so, since it is a new measure, all they really had to do was just face validity, right, with this – with the structure process, right?

Karen Johnson: Right.

Male: To pass.

Karen Johnson: Right.

Male: But, they didn't – they didn't do that. So, they – so, they tried to do empirical testing which – I give them credit for that, right? I mean that's better. We – that's what we want. The problem with this is being, you know, that this only used – is the way they did it.

The methods themselves were fine. It's not a very high bar of validity. But, it is – it's an attempt to show our construct. It's just you are trying to estimate a construct with three or two plans. So, I think they could take their data and think about ways in which they could break up their plans into smaller pieces in some sort of meaningful way and try again with some of the same results.

In some ways, instead of starting at the higher level and estimating – and doing your analysis at the highest level, you almost want to flip it and say, "OK, like what's the smallest unit I can – I can – I can test?"

And if it's reliable and valid at that level, it probably – will work at a higher level. But, I can kind of see how they (went and fell) into this sort approach without really stopping to think about what it is that they were doing and what it actually meant.

Karen Johnson: Yes. I think – NQF does not have thresholds really for how many – how big the sample size has to be. But, I think it seems like, you know, their hypothesis weren't quite met because of the sample size problem. So – OK. Anything else you'd like to mention on the – on validity? OK. Go ahead and cast your vote for measure 3484 on validity.

And, then, for the composite construction, the ratings were four "High" and one "Low." So, unless you want to revote on that, it would pass that criterion with a "High" rating. Again, they used the same methodology in terms of empirical testing. They used the Cronbach's alpha methodology to compare the components.

And in terms of their rating methodology, I don't think they provided empirical analysis, a sensitivity analysis, but did describe their advisory panel input for the decision not to wait anything differently. Again, this is all or none. So, they weren't – they weren't putting more emphasis on Influenza versus Tdap. So, I don't know if – I will let you guys tell me if you want to revote the composite criterion or if you are OK with it going through with a "High" rating.

Male: I suggest – I mean I think to be consistent with our process, we would not vote on it.

Karen Johnson: OK.

Male: But, I don't mind voting if others want to reconsider. But, I didn't think we were supposed to go back and re-challenge.

Karen Johnson: No. In general, not. And I don't think I heard anything from the last conversation that would necessarily make me think you guys might have changed your mind about this one. But, is everybody OK not revoting on this?

Bijan Borah: Well, this is Bijan. Actually, I already kind of highlight this (inaudible) and I don't think I can undo (inaudible). Yes. No, I think I am fine not to revoting.

Karen Johnson: OK. All right. Why don't we not revote on that. That sticks with our process the way we have laid it out. And I think it's consistent with our discussion in the last measure as well. OK. So, we are finished with those two measures.

We did have a couple of other measures that passed. They are listed at the bottom of page one and top of page two in your discussion guide. One is Defect Free Care for AMI and the second one is the In-Hospital Risk Adjusted Rate of Bleeding Events for Patients Undergoing PCI.

Both of those measures did pass. And we are not planning to discuss those measures unless one of you want to pull one or – one or the other of those today. So, we'll give you just a minute. If you want to pull it, we will allow you to pull it right now. Otherwise, we will not be discussing these measures.

David Cella: No nomination here. This is Dave.

Karen Johnson: OK.

J. Matt Austin: This is Matt. (I'm fine with) …

Karen Johnson: OK. All right. Great.

Bijan Borah: (Inaudible). This is Bijan.

Karen Johnson: OK. Good. OK. So, that actually concludes our business for today's call. And we will be in touch with you a little bit later to let you know what the overall voting results, et cetera are for the various measures.

Before we let you go, does anybody have anything you'd like to mention in terms of the process of things that you want us to think about going forward or do you just want the rest of your day back? Either way is fine with me.

Male: Thank you for the setup – for setting everything up and walking us through it, Karen. It's very helpful.

Karen Johnson: OK.

Bijan Borah: And this is Bijan. Actually, I really like the process this time. I mean it was really helpful. It's much better than what we were used to.

Karen Johnson: OK, good. You know, we know there are still some things that we'd like to work on. But, we do think, in general from our side, it seems to work really well too. So – all right. I'm not going to belabor this.

Thank you, guys, so much for attending. We will be in touch with you soon. And we'll talk to you – probably the next time will be on our monthly call, which is scheduled for some time in November, I think the 8th but we don't remember exactly. But, we have the monthly call coming up pretty soon. We'll talk to you then.

Male: Thanks, Karen.

Bijan Borah: Good night. Thank you.

Karen Johnson: Thank you. Goodbye.

Male: Thank you.

Male: Goodbye.

END