NATIONAL QUALITY FORUM

Moderator: N/A
March 19, 2019
3:35 pm CT

(Miranda): Good morning everyone and thank you for dialing into today's Assigned (unintelligible) Methods Panel Subgroup Member 3 Follow-up Call. We'll start off today with some brief housekeeping items. And we'll (unintelligible) before we dive back into the measured discussion.

So for subgroup members, we resent an email that went out last week with your survey monkey link and the discussion guide. So if you'd like to pull up that email from this morning, it should have everything you need for today's call.

Last week we discussed two measures related to the discussion, 2539 and 3508. And today, we'll use this time to discuss the remaining two measures, 3511 and 3513. And we will briefly be revisiting 2539 as well.

As a friendly reminder, this is a public call. Developer representatives are on the line to answer any questions from NQS staff or from the Methods panel members but there are no opportunities for public comment.

So with that we'll do a brief roll call. Do we have (Christie Chieftan) on the line? (Karen) (unintelligible)?

(Karen): (Unintelligible).

(Miranda): Hello. Hi (Karen), thanks for joining. (Susan White).

(Susan): I'm here.

(Miranda): (Ron Walters).

(Ron): I'm here.

(Miranda): (Jan Perlow).

(Jan): I'm here.

(Miranda): Great. And (Jack Needleman).

(Jack): Here

(Miranda): Wonderful. All right. So I am going to turn it over to (Ashley Wilton).

(Ashley): Hi, good morning everyone. Thanks (Miranda) for that and getting us started. So as (Miranda) mentioned, we are going to pick up where we left off with the cost measures - 3511 and 13. We'll dive back into those.

And we are going to revisit briefly 2539. If you recall, that was the colonoscopy readmissions measure. And we just wanted to briefly bring that back, you know, after we debriefed as a team on some of the discussion with the liability, specifically with the liability.

We did want to bring that back and redirect the message panel to revote on the reliability as it is in the submission form, the testing form currently. With all the confusion with the projected three year scores and as we - and kind of all the back and forth.

And as we really kind of stepped back and digested everything that happened, it would require (unintelligible) changes to the submission form beyond what we would allow at this point in the process. So my apologies for, you know, misdirecting in terms of what the Methods Panel should have been voting on.

It was a (misstep) in progress and so we're taking this opportunity to correct that. And so what we'd like to do is to just bring the committee or bring the Methods panel kind of back to that reliability discussion and refocus in on those - that discussion around the one year and three year data and clarify exactly what is in the submission form.

We will give the developers an opportunity to clarify anything just so that we're all starting from Ground Zero on that discussion and then we'll have you revote. We don't need to rehash everything but, you know, we've only got a few minutes for this discussion but just wanted to give everyone a little bit of background on why we're back here.

We don't like to rehash and bring stuff back up unless we absolutely have to. We did want to make sure that we clarified the process for that. So at this point if you want to kind of focus in on page three of your discussion guide, we will focus in on that reliability section again.

And in my review of that measure information form and of the testing form, in terms of the specifications that are presented, there is no detail in the measure information form specifically about the measurement period which I looked for in the denominator details.

But we should be directed by what is in the testing form and what they have submitted in terms of the data that was tested for reliability which appears to be only one year of data from 2017. And then the three years of projected data.

So with that I did want to give the developers and opportunity. I think, I do believe they are on the line to just clarify that and make sure that the Methods Panel understands what was done for the testing as is represented in the testing form.

And we'll give the Methods Panel an opportunity to ask any clarifying questions and have any discussion as needed.

(Craig): Hi, this is (Craig Bresendale). Good morning. Can everyone hear me?

(Miranda): Yes, hi. We can hear you.

(Kurt): I believe you characterized the form correctly. We have one year of data that is represented in there. And then three years that we projected based on assumptions of what the data would look like, if we use three years of data.

And so I won't bring up anything outside of that per the directions but I will say that one year of data does produce reliability estimates that are considered moderate to high in the literature of several different articles that present kind of ranges for these types of things.

There is no general consensus in the literature about, what is Mt. Everest's low versus high? But in general, anything about .6 or around .6 is considered moderate in the literature. And generally it' just up from there to high, in the ranges where it turns the highest varies.

And we still do believe that even though we projected three years of data that even one year of data signifies moderate to high reliability. So I will leave it at that.

(Ashley): Thank you (Craig). So at this point I just wanted to give the Methods Panel an opportunity to ask any questions or offer any comments based on that.

(Miranda): No, no questions.

(Ashley): Would you guys be ready to revote at this point - or, I know we had quite a bit of discussion on the last call. And any questions on process or, you know, anything that you would need to discuss before submitting a vote again for that for this measure on the liability?

(Susan): This is (Susan). I have a quick process question, just on voting in general.

(Ashley): Sure.

(Susan): We're voting on reliability only but in the survey monkey it gives us the opportunity to vote on both so do we just not select anything for validity?

(Ashley): Correct.

(Susan): Okay, thank you. Sorry. Silly process question but I just was curious. Thank you.

(Ashley): No, it's a good one. Thank you.

Man 1: So where do I clarify this run? I would like to clarify - again, I think you were - I think this is what you said but I want to make sure. So the data we're presented is one year data with a three year projection. And the data on one year, we have the numbers we have.

We are voting on the one-year standalone regardless of what the measure says or the three year projection as discussed numerous times. So clarify again exactly which reliability measure we are voting on.

(Ashley): So you're voting on the one year. That was from actual data. We'll vote on the three years of projected data. I know there was some discussion the last call about the comfort label of projected data versus actual data which is kind of what sent us into that other track. So the idea is that you're voting on exactly what's in front of you and it's kind of your evaluation of what the three year projected data would tell you.

But essentially I would just clarify that the one year data is of actual data. And so yes, that's what you're voting on.

Man 1: Got it. Thank you.

(Jack): (Ashley), this is (Jack Ewing) and I'm in China at the moment.

((Crosstalk))

(Ashley): Thank you for your diligence in joining.

(Jack): Can you remind me where the survey monkey link is all the materials we've gotten?

(Ashley): Yes, I believe (Miranda) resent an email this morning so it would be kind of at the top of your email. (Miranda), I'll let you give the details on that.

(Jack): Okay.

(Miranda): Sure. I sent it out maybe ten minutes before the start of this call but (Jack), I can forward it directly to you now.

(Jack): Okay, because the email I've got says survey money link below and I'm not seeing the survey monkey link below. Ah, got it. Never mind. We're good.

(Ashley): Okay, so if we're good on that it sounds like everyone is submitting their votes at this point. Reliability for 2539. And we can move onto the cost measures. So I'll give a few seconds for you guys to finish submitting your votes.

And while you're doing that, and just give a quick recap of where I believe we left off in terms of the cost measure discussion. And then we'll dive back in. So with 35, we finished voting reliability and validity for 3508 which was the episode for elective PCI.

There were some concerns particularly with the low reliability scores at the clinician level or the 10-MTI level. And there was some discussion …

((Crosstalk))

(Ashley): … comments on threshold that should be considered by the panel in terms of the (unintelligible) paper and other literature that's out there. In terms of, you know, the low signal to noise ratio and where that cutoff might potentially be in terms of the Methods Panel's evaluation's (unintelligible) measures.

And then there was a discussion on validity as well. The developer did submit information for both face validity as well as empirical validity. We had some extensive discussion on NQF requirements for face validity and we clarified that.

What was submitted by the developers for face validity did not meet NQF requirements for face validity. It was not a systematic evaluation of the validity of the measure which put the focus of the discussion in terms of validity on the empirical validity.

And the committee had - I'm sorry. The methods panel had a discussion on that as well and expressed some concerns about that approach and whether or not it actually demonstrated validity and whether the kind of association with the other measures of high resources or costs with - in some sense kind of validating the measure against itself.

And so with that I'll just kind of pause and see if there's any other kind of summary comments based on our review last week. And mainly because I think some of those issues will carry into our evaluation of the last few measures.

And so to the extent that we want to kind of highlight any of those overarching issues now would be a good time and so we can kind of dive into each of the measures individually. So I'll pause there and open it up to any others for comments before we dive into 3511.

(Jen): The one thing I would point out just for ourselves to be careful about is the difference between the performance and behavior of a procedure episode. So when there's a surgery the variance should be tighter and it should - it's a bit more sort of well behaved.

We have a couple of condition episodes. I think Pneumonia is one of the ones. Yes, that's 3513. So anyway, just, you know, they are, well behave very differently when you look at different types of episodes. Just something to keep in mind as you're reviewing the results.

And I think the developer has added in an extra (inulvative) source for some of the condition episodes so they are comparing to - I'm not remembering off the top of my head. But there's a third sort of reliability - I mean validity piece that comes into play by comparing to an outside data source.

So that's just one nuance. Not all the measures got that. Let me look up the data sources.

(Ashley): Thanks (Jen). I appreciate that and I'll do the same. You're right. There were some nuances with the validity episodes.

(Jen): Oh yes. It was Pneumonia, specifically, yes. And I'll look up the data sources.

(Ashley): Okay. So while she's doing that let me just give a quick overview of 3511. It is a measure of episode based cost for (unintelligible), chronic critical (unintelligible). It also includes services based on the - that are clinically related to the attributing clinician's role of managing care due to 30 days prior to the event through 90 days after the trigger.

Beneficiaries who were eligible, enrolled for Medicare Parts A and B are included. And so for this measure, it's stratified into five subgroups, based on the location and complexity of the procedure. So intravascular, above (knee). Intravascular, below (knee). Intravascular, above and below (knee). Open, above (knee). And open. And then below (knee).

And so again, they have an explanation of why they selected these various stratification groups there for you. They did risk deduction again with (ACCs) but there are no social risk factors included based on the results of their empirical analysis.

And for reliability which is where we'll start, there was consensus not reached based on the preliminary analysis by Methods panel subgroup. There were three moderate votes and three low votes. And again, there were concerns with the movement between quintiles as well as the kind of threshold for low reliability, signal for noise scores after ten and PI level.

And let's see. There were also, in terms of the concerns with the significant movement between quintiles, again, concerns with the application of this measure for value based purchasing and how that might affect payments or rewards for - within programs.

And so I'll pause there and see if there are any other additions to that list in terms of concerns for the reliability for this measure and for other comments from the Methods panel.

(Jack): Okay, so this is (Jack). I guess I will start. I'm looking for the exact reliability from the signal for noise estimate here.

Man 1: 741 I think.

(Jack): And the reliability in the Pin on average is about that .7 that (Adam's) tested. And (unintelligible) low/not low, it was 20% in this classification. The standard for the Pin API falls below that which suggests there was even more misclassification. (Test/retest) correlation is close to the level - I recollection - my recollection of the literature here, I can't say I've gone back and checked is text/retest correlation of .8, sort of at least one source. I have seen the split between you can use for individual classification and when you can't.

And this one falls below that. The, if you look at the testing form and you look at the impact of this in the quintiles as sort of the evidence of, you know, what actually, what level of misclassification one can see at the Pin API level, only half, less than half of the group that's classified in the lowest quintile in one sample is classified in the lowest quintile in the second.

And about a quarter are classified in the extreme, in the top three quintile away from that on the - for those that were classified in the highest quintile, 57% wind up being classified in the highest quintile with roughly 18, 20% in that lowest three.

This one, if you go to just the Pin level, the numbers are somewhat comparable. Perhaps a shade better but not much. To me, this one is right on the border of where I would say - I don't expect perfection in these measures. There's got to be a certain level of misclassification.

But this one is on the border and on the wrong side of the border to me for enough reliability to be used where people are getting paid.

(Ashley): Right.

(Jack): Misclassification of, you know, what looks like 2, 3, 4, 5 deciles very likely and on a ten decile distribution score for payment. That just feels a little high.

(Ron): This is (Ron). I understand everything you said. That's why I asked the question on Friday about, how literal we should be about the (Adam's) classification.

And then we talked about previously purpose of the measure. And it is very hard to divorce the measure from the purpose of the measure. And I think that's a separate discussion.

I don't understand the variation in quintile. I think that's a very valid point that on test/re-test you should not dramatically switch your quintiles and have a high reliability so I will ponder that and consider switching from moderate to low. Consider it.

(Ashley): Hi this is (Ashley) from NQF. I did just want to - based on (Ron's) comment about the use of the measure. I think, you know, the measure is not - will not necessarily be endorsed specifically for use in the MIFs program although we do have some contact and background information that this measure may be used there.

But endorsement is for accountability purposes so for that broad kind of category of accountability, I think that should be a consideration in terms of how the measure can be used and for what type of accountability program., You know, it's not being specifically endorsed for MIFs but for accountability.

(Ron): And that's why I - this is (Ron). And that's why I worded it the way I did, is it's almost like - that again goes on my list of feedback to give to the steering committees that - almost exactly what (Jack) said.

This is a borderline measure for accountability and if it's used in a payment program, serious reconsideration should be given. But I don't know, you know, when measures come to the risk measure, will go to the clinicians workgroup I'm sure.

How does the workflow accommodate that emphasis on utilization and purpose? And I don't, I mean I don't know how to do that other than just in the recommendations. As (Jack) said, it's right on the border there for just plain old measurement. It's not bad.

(Ashley): Right. I would again just you know, encourage you guys to kind of consider your readings for reliability and validity in the context of accountability which I think is fair game in terms of kind of process going forward, if the measure makes it past the Methods panel and the clinical test.

The committee will have an opportunity to kind of weigh in on a whole in the measure and they will be applying the use and usability criteria in which that will be - there will be some low level of discussion there as well. So hopefully that's helpful.

Woman 1: But it does seem like it pushes towards the conservative estimation, right. It can be used in a broad range of ways. In some sense we have to think about payment, for example, the most sort of risk use or the most impactful use.as the minimum in our minds.

If - anyway, that's how I would solve that dilemma.

(Tray): This is (Tray Jarvin) from Acumen. Could I actually offer a couple clarifications from the discussion so far?

(Ashley): Sure. Go ahead (Tray).

(Tray): Thank you. The first thing I wanted to make sure to note was for the quintile shift analysis, we briefly discussed this last time but this, for the procedural measures, is using one year of data and doing the split half testing.

The implication of that is that if there's a provider that in a given calendar year for the purposes of a value based program has 60 episodes, here in this table it's almost as if, we're looking at 30 episodes for them.

And so in the context of this discussion about how the measures are going to be used, I think the quintile rank shift tables are not at all indicative of how the measures will actually be used. Because in essence you can think about it as taking a provider who has X number of episodes and the precision that comes with that and then dividing that by two in terms of the number of episodes and the implications of that for precision.

So I would say that the precision rank shift table is not giving you an indication of how this would actually be used in the measure and in any program. The second clarification is about the correlation. We just submitted the correlation information as a supplement to the more conventional reliability metric with the comparison between variance and between costs between variance that you all saw.

And so we just submitted the Pearson correlations I should not that often measure developers when they're relying solely on that metric will use an intra-class correlation coefficient, typically ICC21 from (unintelligible).

And the calculation of that would differ from the Pearson correlations in general. And so I just want to make sure that the Pearson correlation isn't being interpreted in a way that is comparable to what other measure developers may submit in terms of ICC21 for those sorts of comparisons across time.

And then the final point is, in terms of how these measures are actually going to be used, (PMS) in the past has considered for measures whether both a Pin and Pin MPI version should be available. And has made decisions in the past to make only a Pin version as a measure available for use in a value based purchasing program like MIPs.

And so I would encourage that if consideration is about the use of these measures in accountability for (MIPs) in a value based purchasing program that the fact of thinking about a Pin and Pin MPI measure as separate is extremely useful because in practice those considerations are taken into account by CMS. I'll stop there.

(Susan): So this is (Susan). So I have a couple of comments on that. So whether it's Pin or Pin MPI is irrelevant. Statistic is the minimum sample size that would be allowed. And so, before someone would be eligible for measurement.

And I can appreciate the impact of a split half in the quintiles and what I would encourage the measure developer and other measure developers to do is to present the data. So if this represents a floor of 390 instead of a floor of 60 for instance and that's not the case here but just as an example, to explicitly express that would be very helpful for the committee.

But I don't think we have the option of sort of bifurcating this into approving Pin, not approving Pin MPI, something like that. I think we have to take that measure as submitted is my understanding. But maybe I should state that as a question and not a statement.

(Ashley): Yes. (Susan), this is (Ashley). That's correct. And I just wanted to point out that the measure submitted is submitted for two levels of analysis which is both individual clinician which should be the Pin MPI and then the clinician group.

So both of those should be taken into consideration based on the measure as specified.

(Tray): And (Susan), one quick note. For the case minimum question, the implication that we have in mind is that for the Pin level of the measures, the distribution of number of cases will be substantially different than Pin MPIs.

And so that's what we have in mind. And in the case where maybe they're solo practitioner Pins that have a small number of cases that sort of function like Pin MPIs. And like I was saying, they'll be a substantial portion and the distribution lies with sort of a higher, systematically higher number of cases that the Pin MPI measure.

(Ashley): Right. And so the minimum for the measure should potentially be higher than, I don't know if this one is 10 to 20.

(Susan): This one's 10.

(Ashley): Yes, and so what I think what we're saying is one way this measure can overall be strengthened is by having a higher minimum case requirement.

(Susan): Exactly.

((Crosstalk))

(Susan): The Pin MPI is kind of relevant to us.

(Ashley): Right.

(Susan): What's more relevant, I think, at least in my mind, is what you're setting as the floor. How you get to that floor is in implementation, so.

(Ron): Yes, this is (Ron). I like how this has evolved over time whereas we went from just wanting to see the deciles and many of the measures that they've gone on given the (unintelligible) information distribution.

But I think now we're on another matter which is very important and (Jack) brought it up, is if the volumes are low enough that they affect the test/re-test, the measure developer should provide that information and/or have a larger number, a higher minimum for the test/retest.

I really would like to see that because then it would take care of the discussion that we're having right now.

((Crosstalk))

(Jeff): Yes, we have - (Jeff) again. And again, I'm hearing on the quintiles table. And while the sample size is maybe a larger Pin and the Pin MPI, distributions are awfully similar which would suggest that it may just be the reliability of the measure.

But it would not be serving to go through all this documentation. I'm told the documentation doesn't really tell you what you think it tells you about reliability because that's what we're trying to make sense of here.

If the issue is that the - again, I think that the committee is actually going to have to go back and make some decisions as to whether they're prepared to - whether we're prepared (unintelligible), whether we're prepared to standardize some recommendations here based upon what we've seen as we cross a number of different things.

I'm happy to apply a committee adopted set of standards here for reliability. Right now we're moving our way towards that and we each have our own judgement. And as we talk to one another, (collective) judgement.

But if what the developer is saying is, you know, you've only got half the cases that you would see in the natural world here, I'd be just as happy seeing test/re-test to, you know, sample with replacement models of the appropriate size for each unit, we'd get the same kind of distribution of variance as we get randomly splitting. And we'd get the right sample sizes for the measures. But that's not what we've got in front of us and that's not the basis on which we can decide. And what we've got to decide on is the data that's been provided to us. And this doesn't look like a terribly stable measure.

(Ashley): Right. Is the - I don’t want to push too soon, but are you guys - do you feel like you’re at a point where you would be ready to vote on reliability at this point for this measure?

(Susan): Yes.

(Sam): This is (Sam), yes.

(Ashley): Okay. Let’s go ahead and do that. And then while you’re doing that, I would like to just revisit again this issue for validity for this measure which was scored at moderate but in looking at the submission form again in response to (Jen)’s initial comment on kind of the nuances between the measures, I was double-checking to see what additional validity, empirical validity testing was submitted.

And I believe for this measure it was very similar to the measure 3508 in terms of the face validity and then the empirical validity with the comparison of the ratio we observed of expected spending.

So I wanted to see whether or not the discussion for the last measure would warrant a re-vote for validity for this measure as well. Can we hear some thoughts on that?

(Susan): I think the answer is yes because of the face validity clarification we got last time. That that may have driven many of us and would have been a significant portion of our rating. And that’s now sort of off the table so it changes the equation for me.

(Ron): Well it forces, this it (Ron). It forces you to use the empirical validity as being sufficient enough that you don't - you're not concerned with the face validity. That's what it does. That was the discussion we had.

(Ashley): Yes. Well said.

(Sharif): And this is (Sharif) from Acumen. I know a lot of time there's the discussion of the empirical validity and the use of the Pins based measures to examine correlation and stratification.

The thing that I want to emphasize is that we're not testing the data element validity by using Pins based measures so we definitely appreciate that you wouldn't want to test the reliability or validity of Pins data elements using Pins data.

Instead what we're doing is testing the empirical validity and sort of the construct validities of the measure. And we reviewed past and endorsed measures and a series of claims based NQF endorsed measures. All used comparisons to Pins based metrics.

And so examples are the total cost of care and total resource use measures. And I would, you know, earlier a point was made about examining the correlation with an external measure with MSPD. I think that examination is exactly analogous to the sorts of correlations that we're showing in a testing submission for the revascularization measure.

The reason I say that is that the MSPD measure itself is a claims based measure. In fact it will include the exact same claims as other measures, other cost measures.

And so I think what we'd like to focus on in the empirical validity section is really thinking about the question of clinically, if there are certain high cost events that we can identify whether through identifying readmissions or complications and so on versus events that are more normal versus care. Is the measure actually requesting this?

And for the revascularization measure, we assume the correlations that you see. But the correlations with MSPD or any other metric, even if you had a correlation with a readmission metric, they're claims based metrics that are examining the same sort of concept.

And just as a quick supplement to what was submitted in the testing forum, Dr. (Sanders) here will walk through for revascularization additional correlations along this line very quickly that could be useful for revascularization.

Dr. (Sanders): Thank you (Sahrif). So with revascularization, as (Sharif) was mentioning, in addition to showing that this has a downstream readmission and post-acute care associated with higher costs, we looked at different clinical categories of costs and showed that a number of complications that are complications or additional costs that clinically we know are important with regards to efficiency were also associated with clinicians' performance.

So amputation related services, complications related to bleeding, renal injury infection, repeat revascularization, room care services, those are four different categories which fairly consistently increase as we go from clinicians who were high performing on the measure with low costs to clinicians that were low performing with high costs.

So even though that the largest absolute driver of costs was the cost of the initial revascularization, all of these other areas that we know are important drivers of clinical care and costs for this procedure were all associated - were all higher in the worse performing clinicians.

And I think that really does support the construct validity. And similar to what we saw with other measures we also looked at that same correlation of use complications with the clinicians' average patient risk and we saw that those are also positively correlated.

So in our risk model, clinicians taking care of patients that we found were more sick also ended up having higher costs related to these different complications, I think is also supported.

And kind of a third point related to that to just briefly bring up is that we compared the reliability of our measures to a very similar measure that didn't include any service assignment.

So a measure that's very similarly constructed to (prior) endorsed episode cost measures that included all costs over a given window. And we almost universally found that the service assignment and the inclusion of costs that our 22 clinical experts selected specifically through revascularization improved the reliability of the measure at both the Pin and the Pin MPI level. Thank you.

(Jen): So this is (Jen). Two quick reactions to that. One, I would love a world where you could validate claims with claims. I call it validity light. And trust me, I'm guilty of torturing the claims data as much as possible for validation purposes.

But the hypothesis test that would impress me the most would be to pick a sequelae or complication that we know is possible but is actually a low cost event and to look at the difference between episodes, cases that had known low cost sequelae and known high cost sequelae.

So you're always showing us the high costing since we know that it costs more. But, you know, our more nuance measures are able to differentiate between the high cost and the low cost event, so just as a thought or comment based on my own experience working in the episode space. But again, all of what you just described is very helpful.

Dr. (Sanders): Just as a quick response to that, we've identified for revascularization specifically one (example) here to your point.

(Jen): Sure.

Man 1: So for example, a low cost clinician with an average cost of $16,500 compared to a high cost clinician with an average cost of $25,800. Room care services are definitely increased in the high cost clinician but definitely not driving the distance, being only $400 in low cost and $800 in high cost.

And similarly, there's rare neurologic and bleeding complications which are $350 in the low cost clinicians and linear increases to $460 in the high cost clinicians.

So these are complications or services that are definitely not driving the overall episode cost but are correlated or associated with the performance of costs in other areas.

(Jen): Right. But I'm interested in the case level, not the provider level, right?

Dr. (Sanders): Yes, and I'm actually now confused because the documentation says that these correlations and the looking at, expected to observe, where there were episodes with (facts) - specifically speaks to episodes not clinicians.

(Jen): Right. Okay.

Dr. (Sanders): And the last couple of discussions for this, you've talked about it as if it's a comparison of clinicians, high cost clinicians versus lower cost clinicians and then high cost episodes versus low cost episodes.

(Jack): Yes, we wanted to provide supplementary evidence based on people's interest and thinking about measures as used in the program. And so the numbers that Dr. (Sanders) provided focus on the measured scores because of that interest.

Now again, the documentation says you're looking at what's expected at the episode level. A few months ago you were talking about what was expected at the clinician level averaging across all the clinicians' cases.

((Crosstalk)

(Jack): But that's what the application says.

(Ashley): Right. This is (Ashley). I think this is where kind of confusion erupts oftentimes is that, you know, the Methods panel has spent time evaluating and looking at the materials as submitted. And while I think its helpful contact sometimes to hear other analyses that's been done, it's very difficult to analyze and discuss information on the fly like this which is why we allow kind of this additional information at this point in the process.

So I just want to refocus the committee or the panel on what’ in the submission form because that is where the basis of the vote should be focused. So, you know, I don’t want to kind of allow kind of a conversation kin that it’s going to introduce more confusion into the actual basis of the evaluation so (Jack), your characterization is correct.

(Jack): Thank you.

(Ron): This is (Ron). For the white paper writers out there, these issues keep coming up that should be addressed as far as for the future as far as the claims versus claim validation and episode provider relationship for measures anyway. That’s a different issue.

(Ashley): Those are good point though.

(Ron): They are.

Man 1: And if it would be helpful (Ashley), we could send the sort of information that we just discussed for consideration. I know that may not be in line with the procedure but we wanted to offer that because we think it is important.

(Ashley) So again, at this point in the process, I mean if folks are interested in seeing that information as kind of informational. It wouldn’t be used in terms of evaluation of the measure at this point. And so we can maybe discuss offline how that would be most useful.

(Frank): I have all I need to vote. This is (Frank).

(Ashley): Are others ready to vote as well?

Woman 1: Yes.

(Ashley): Okay.

Man 1: Yes.

(Ashley): Okay. Thanks. Go ahead and vote and let me - give me a second to queue up the next measure which will be 3513 for simple pneumonia with hospitalization.

So this measure simple pneumonia with hospitalization cost measure looks at the clinicians' risk deducted costs (offered to) beneficiaries who receive treatment for this inpatient episode.

And again, the measure is based on a trigger for the attributed clinician's role of advantage in the care for the 30 days prior to the clinical event and through the hospitalization. And includes the Medicare beneficiaries enrolled in the Medicare Parts A and B in the performance period.

Again, this measure is also stratified into three subgroups - simple pneumonia with hospitalization without complication of comorbidity since simple pneumonia with hospitalization with complications of comorbidity with simple pneumonia hospitalization with major complications or comorbidity.

And again, there's details around how those stratified subgroups are defined and triggered. This measure for reliability actually receives a low rating and so we won't rehash that. But we do need to adjudicate the consensus not reached for the validity.

The ratings for validity were one high, two moderate, three low and zero insufficient. Again, consensus not reached.

Man 1: Low reliability.

(Ashley): Score? Yes.

Man 1: (That's lower than the rating for validity, I'll tell you that).

(Ashley): Yes, exactly. We fail, right. Yes, so for - again, so similar issues again but this measure was one (Surgen) commented on earlier that there was similar information submitted earlier along the lines of comparing the pneumonia measure scores to the Medicare spending preventing measure.

And so again, there was face validity which did not meet NQF requirements. The similar validity, empirical validity testing around correlation with hospital admissions in post-acute care looking at the observed to expected costs. And then also again this comparison to the Medicare spending prevention measure.

So with that I will open it up to the Methods panel for discussion (to discuss) this measure.

(Susan): I think this has all the same issues as 3511. My bigger question - this is (Susan) by the way. Sorry. My fear question, I don't even want to ask is that now that we know that face validity doesn't - assuming that the measure developer did the same validity testing for all the cost metrics.

And we're only discussing a few of them so I don't know how we - if or how we should address that with the other metrics because we have sort of this new information on face validity.

You know, I don't want to open a can of worms but I'm going to lay it out there.

Man 1: You did.

(Susan): Sorry.

(Ashley): Yes, good point. This is (Ashley). So the plan is to, you know, kind of get through these two measures and then as is the process with all of our calls, we have the opportunity to pull measures either at staff or Methods panel discretion that were not identified for discussion.

So that would include measures that, you know, pass or didn't pass or what have you. So there will be an opportunity to do that. And so we can revisit those three measures and decide whether or not there needs to be a reconsideration of the validity for those as well.

(Susan): Thank you.

Man 1: Thanks (Susan).

(Susan): Happy to help.

(Ashley): So, just to kind of queue this up again, is there any additional discussion that you feel is needed for validity on this measure based on the, kind of the different, the additional kind of type of validity testing that was submitted for the Medicare spending, Medicare spending for (unintelligible) measure or feel like this is analogous to - close enough to the other empirical testing that it all kind of goes together?

(Susan): I will - the developer was pointing out that the MSPB measure is (claim) space. It's a broader spending measure so it's not like you went and got hospital or delivery system cost data. It's not an external source.

Man 1: And I should note that if we had sort of hospital or delivery system cost data, that would be measuring a different concept than the concept that CMS is interested in.

When we say costs here, we're thinking standardized allotment amounts so this is costs for the Medicare Trust Fund including beneficiary copayments. It could differ very systematically just based on our own analyses of charges and cost reports at the hospital level from any certain costs that are reported by hospitals.

So that's the reason that we avoided any sort of comparison to hospital charges or costs.

(Ashley): Okay, amended claims data. I got you though. So are you - do you feel ready to vote on validity at this point for 3513?

(Susan): I am.

Man 1: Yes.

Woman 1: Yes.

(Ashley): Okay. Go ahead and submit your vote and let's kind of regroup here. There were three other cost measures that passed with either high or moderate reliability and validity scores.

The cataract 3509, cataract removal with intraocular lens implantation. 3510 was the screening surveillance colonoscopy. And 3512 was knee arthroscopy. So we co within this discussion guide have a summary of those measures. I can tell you what page here. Just a second. If someone gets to it sooner, just let us know.

Okay, page 15 has measure 3509 which is the cataract removal and lens implantation measure. And so I think just to make sure I'm hearing the panel here, it seems like the concern - overarching concern here is validity, correct?

(Susan): Yes.

(Ashley): Okay. And so what we can do if the panel is in agreement with this is to have an opportunity for discussion, if you feel it's needed and then a revote on validity for these three measures. Does that sound accurate?

Again to recap, there isn't the detail in these summaries that we had in the measures that were for discussion because they passed. But essentially there was, you know, the same testing submitted in front of - the same information submitted in terms of testing for validity for each of these measures.

So again, the discussion is - do we need additional discussion for validity at the individual measure level for each of these or would you be ready to vote?

(Susan): Since the method was consistent, I'm not sure there's - I don't know about everyone else but I think I'm ready to vote.

(Karen): So you're just opening these up for revote in case people think that the misinterpretation of face validity was driving their vote?

(Ashley): I'm sorry. Who's speaking?

(Karen): I'm sorry. This is (Karen). So you're opening it up for revote to make sure that we all understood the face validity issue.

(Ashley): Yes, well I think the concern is not - yes, the concern was that the votes for those measures were also submitted based on the understanding that face validity was adequate.

And given that that is now off the table, the consideration and the discussion that has happened around the other methods of empirical validity testing that was submitted, there were several concerns raised about that.

And that would give - by re-voting on validity, it gives folks an opportunity to capture votes based on the criteria and what was actually submitted for empirical validity.

(Karen): Got it.

(Ashley): I don't see any reason for continuing to discuss unless someone thinks there's a different mechanism within these measures than we've already covered?

(Karen): I'm fine with that.

(Ashley): Okay. So let me have you guys go ahead - just on validity and to the point earlier, you don't need to revote on reliability for each of these measures. It would just be the validity criteria and for 3509, 3510 and 3512.

Okay, so with that done, I think that completes our evaluation for these measures. So we will collate all of the votes and we will follow-up with the Methods panel as well as the developers. With final votes in, next steps, we will have an opportunity to review that and summarize the votes for all the measures that were discussed by the subgroup.

Any final thoughts for either recommendations for the developer going forward or questions about next steps? Okay hearing none, I just have one other …

(Jack): Oh wait. Wait. I'm sorry. This is (Jack).

(Ashley): Okay. Go ahead.

(Jack): First in 9, 10 and 12, we're voting on validity but I just want to say one thing which I said earlier in this process about the work that Acumen has done for CMS here which is I think they've actually done a very good job ln this.

The work to build consensus in the provider community about what should be counted and what shouldn't is commendable. And if we're not basing things on just face validity, they have looked for ways to provide empirical comfort if the measure's doing what it should do which is measure high use of resources.

The reliability issue has been largely one of degree of variance within the individual clinicians and partly driven by sample size but also driven by inherent variability. While they have been aggressive in serving CMS to try to argue that virtually all levels for tests with reliability meet the standards, we don't necessarily share that.

The actual analysis of the reliability has provided a high level of information for us. We suggested ways to improve it but they've done a good job and I don't think the votes for or against endorsement in terms of reliability or validity are intended to (unintelligible) the work that they've done.

(Susan): Agreed.

(Ashley): And I think that's such a great point. And I want to point out that this might be a somewhat impossible task, right, this notion of resource and episode based resource use measure is fully valid and reliable on its own.

It may not be scientifically possible so we can't blame sort of the developers for that potential underlying problem. You may need to group those to have composites for example. So anyway, just to that point, I do think that what was presented was wonderful in revealing exactly what was here.

(Karen): This is (Karen). I would agree with that. I particularly appreciated the opportunity to be able to see what was exactly going into the model and be able to look into exactly what the coefficients were.

On things like dual status, for example we didn't talk about post-economic status because it wasn't in the discussion guide but that's obviously what I always look at.

I very much appreciated all the information. I actually think these measures with the exception of some of the ones that didn't pass that the reliability is quite low.

I found them to be quite thoughtfully constructed and for many of them I'm not changing my vote of moderate (even in the absence of) the face validity testing because I did think that the other claims based measures is reasonable.

I mean in this case we know that what they're measuring is the claims. So correlating it with something else to me doesn't actually change what they're measuring is fundamentally the claims. But I think that's reflective of quality I think is an issue for the content committees.

I'm not sure that I think all these cases is a reflection of quality but I echo everyone's sentiments in saying that the amount of information presented and the way in which it was presented I think is terrific.

(Ashley): Thank you all for that feedback and hopefully I think there were a lot of really great suggestions from the Methods panel members as well as the developers. So I appreciate your input throughout this process and hopefully the developers found that helpful.

And we'll continue to work with them going forward for the additional measures' submission. So thank you for that. I did also just want to throw out a quick request for the Methods panel members.

I've been jotting down some of the higher level issues that came up in these discussions that may be - we will want to tee up for discussion with the larger group.

If there are things that you know, that I know, that have gotten kind of a clean versus clean validation, kind of a reliability threshold as a couple of the points.

But if there are other kind of broad issues that you kind of identified or came up during this evaluation, if you could just shoot us a quick email bulleted, you know a few words kind of bulleting out what the issue was, I think that will be super helpful.

We're preparing for the in person meeting as well as I think one webinar before that in person meeting. I think it would help us kind of have a sense of what - how we might address some of these larger issues going forward and seeing what guidance we can put forward, that the full Methods panel can put together going forward for developers.

So if you have time to do that today, that would be awesome. And once again, just want to thank you for your time and your thoroughness in reviewing these measures.

We understand they are - there's a lot of material. It is complex information and you guys did a great job working through it so thank you very much. If you have any other questions or concerns that arise, feel free to reach out to us and we will be in touch.

And thank you. Thank you for juggling like 800 things at once.

((Crosstalk))

(Ashley): Thanks guys.

Man 1: Thank you.

(Ashley): Have a good day. We're done an hour early. All right.

Woman 1: Yay. Bye.

(Ashley): Bye.

END