NATIONAL QUALITY FORUM

Moderator: Scientific Methods Panel
October 12, 2018
2:00 p.m. ET

OPERATOR: This is conference # 8665889.

Operator: Welcome everyone. The webcast is about to begin. Please note today's call is being recorded. Please standby.

Miranda Kuwahara: Good afternoon and welcome to the Methods Panel Subgroup Number 3 Measure Valuation call. My name is Miranda Kuwahara with the National Quality Forum, and I'm joined by my colleagues, Karen Johnson, Ashlie Wilbon, Poonam Bal and May Nacion.

We'll begin with a roll call of the Methods Panel Subgroup Number 3 members. We have David Cella on the line.

David Cella: Yes, speaking.

Miranda Kuwahara: Great. Bijan Borah?

Bijan Borah: Yes, I'm here.

Miranda Kuwahara: Wonderful. Matt Austin?

Matt Austin: Good afternoon.

Miranda Kuwahara: Jeffrey Geppert?

Jeffrey Geppert: I'm here.

Miranda Kuwahara: Mike Soto? And Lacy Fabian?

Lacy Fabian: Here.

Miranda Kuwahara: Great. So before we dive in to measure discussions, I wanted to make a few housekeeping remarks. A discussion guide was sent to Methods Panel Subgroup Number 3 members' yesterday afternoon, and that document will guide today's measure discussions. And we'll follow the order presented on that document. Consensus was not reached for the first eight measures presented on the discussion guide and are subsequently slated for a discussion today.

All other measures received passing rating and will not be discussed during today's call unless a member of the subgroup would like to take this opportunity to pull one of those measures for discussion. If you choose not to discuss those additional measures, the decision from your preliminary analyses will be made final for those passing measures. So I'll pause here briefly to give you an opportunity to review those two measures.

And those measures are Number 2377, Defect Care for AMI, and 24 – I'm sorry, 2459, In-Hospital Risk Adjusted Rate of Bleeding Events for Patients Undergoing PCI. OK? Hearing none, we'll continue on.

In that same email containing your discussion guide with a link to a SurveyMonkey, we ask that you pull that survey up now and cast your votes as we move along through the measure discussion. Staff will prompt you when to cast those votes at the end of each measure discussion.

Timing is limited on today's call. We have roughly 13 minutes to discuss each measure and although we would like to come to consensus on all of these measures, we do have a follow-up call schedule on Thursday, October 18th, from 2 to 4 p.m. to discuss any items we don't get to today.

And finally, I do want to note that this is a public call. However, there will be no opportunity – excuse me – for public comment, and subgroup members cannot direct questions to developers. For recordkeeping purposes, we ask that you say your name each time you provide remarks.

And with that, I'll pass the ball over to Karen Johnson.

Matt Austin: Miranda, (inaudible) Matt on. Question.

Miranda Kuwahara: Oh. Hi, Matt. Go ahead. Sure.

Matt Austin: You had mentioned a SurveyMonkey poll. Where do find that? That wasn't clear to me.

Miranda Kuwahara: Sure. So that is in the email that went out yesterday afternoon that also held the material for today's call. It holds the discussion guide and the agenda.

Matt Austin: (Inaudible).

Miranda Kuwahara: And the links you see in the body of the email.

Matt Austin: A link to the committee SharePoint site, but that's the only – oh, there, I see it. OK. It was just small. OK, I found it. Thank you.

Miranda Kuwahara: OK, thanks.

Karen Johnson: All right. So this is Karen. Thank you again for joining us today. We hope this call goes smoothly. We'll see. We've done two so far and both have gone well. We were able to not only come to consensus on the various measures that we needed to discuss, but also I think people, in general, learn from each other. We learned some things, so so far things have been going pretty well. So we hope that you like this kind of change to our process in terms of how we are evaluating and rating measures. And we'd love, after the call, any feedback you might have in terms of how things went today, the materials that we provided you, that sort of thing.

Just a couple of things to reiterate this discussion guide, we wanted a tool to help us facilitate the call so we don't get kind of way off base and off-tangent, talking about fun stuff that maybe not absolutely critical things, so that's why we did a discussion guide. But I will point out that in the preliminary analyses that you guys did, many of you made a lot of comments. Not everyone of

those found their way onto this discussion guide. Sometimes we try to group things and kind of organize things that way. Sometimes just some things felt like they were things you're willing to point out, but not necessarily things that we need to talk about.

So just want to let you know, if you feel like something that you really want to note hasn't made it to the question guide, feel free to bring up any of those kinds of things. That's absolutely fine.

Just a reminder, in terms of what will be presented to the standing committee, this was a question that came up from one of the subgroups, so it was probably a (inaudible). It depends on whether a measure is passed or not. So if a measure is passed so is viability and validity for you guys, then that measure go to the appropriate standing committees.

And along with the measure, we'll go a summary that we will write based on your preliminary analyses and today's discussion, as well as the preliminary analysis forms that you guys filled out. So all of that stuff will go to the standing committee.

If a measure does not pass today or if we needed on a second call, then that measure doesn't actually go in its entirety to the standing committees, so the standing committee will not be considering at this cycle. What we will tell them is that measure XYZ was considered by the Methods Panel. It didn't pass, and we'll provide a very brief rationale for why it didn't pass. But we would not be sharing the summaries and we would not be sharing your preliminary analyses. Instead that information would go back to the developer in hopes that that would help them refurbish and hopefully soon that again next cycle. So that's the process there.

Just another FYI about the SurveyMonkey that you've opened up, that is not a live vote or at least let me put it this way, we can't see your votes right now so we won't know if you guys have actually passed or not passed on one of our criteria. So we will have all the discussions that we need to have realizing that the actual determination of the ratings have to wait until after the call when we get all your – all of your votes in.

I think that is all that I wanted to say kind of in addition to Miranda's opening comments. So with that, just to let you know how we set this up here, we just kind of farmed out the measures amongst our team. So we have different people here at NQF that are going to lead us lead us through the evaluation. So we'll start out by just telling you a little bit about the measure itself and introducing the measure and where the ratings fell. And then we'll segue right into the actual items that we need to (discuss). OK.

So, did anybody have any questions before we get going?

Jeffrey Geppert: Karen, this is Jeff. Can you just remind me sort of the definition of consensus?

Karen Johnson: That's a really good question. It is – we really wanted it to be greater than 60 percent we feel like a measure passes, OK. That's what NQF's rules have always been. We have to think about that a little bit with these groups because of the math involved. We have a couple of measures that have come through so far that they're still going to split a 3-5 split, and that's right at that 60 percent point. Usually, we want more than 60 percent. So we have to figure that out and we have to figure out is that close enough to go ahead and send on the – to the standing committee or is that something that we feel like it's not quite there and will push back.

Obviously, things that are kind of majority, if something goes for moderate, one insufficient, then we would rate it moderate. So that makes sense?

Jeffrey Geppert: Yes, OK.

Karen Johnson: A little wishy-washy, yes. Little wishy-washy we'll get it figured out in the next few days. And quite frankly to tell you the truth, we're going to wait until all of the results come in and make our determination then.

Jeffrey Geppert: OK.

Karen Johnson: Any other questions? All right. Let's get started.

I know May is going to do a first one. So, May …

May Nacion: All right, sure.

Karen Johnson: … walk us through 3309.

May Nacion: Sure. I'll tee up the information here for a discussion. So we're going to first discuss 3309, the Risk-Standardized Survival Rate for In-Hospital Cardiac Arrest. So this is a new measure. However, I'm sure a lot of you remember that this was actually submitted for the Methods Panel to review in Fall 2017. It went through to us and went to the standing committee. However, it was withdrawn from standing committee evaluation due to data discrepancies and their submission materials. So they're back again and to update their information.

So, this measure is an estimate of the hospital-level risk-standardized survival rate for patients aged 18 and older who experienced an in-hospital cardiac arrest. It is specified for hospitals of 20 or more cases of in-hospital cardiac arrest during the measurement period. And the risk-standardized survival rate is calculated by the weighted average on adjusted hospital survival rate for the entire study sample times the hospital's predicted survival rate divided by the expected survival rate.

So this is an outcome measure. The source is registry to get with the guidelines data registry. It is at the facility level for level of analysis. They did perform risk adjustment with nine risk factors.

So for weighting for reliability, I think across the board, everybody was happy with it, so moderate across the board. They performed a score-level reliability testing using signal-to-noise – the signal-to-noise method, and they did not perform data element reliability testing and that's OK. It is a new measure. It's not – it's not needed for this.

Panelists did expect a little bit more desire for more information and just means median of the results. But overall reliability, nobody had any problems with it.

For validity, they performed a face validity, and that is OK because it is a new measure. So face validity is OK. It meets NQF requirement because it is a

new measure. They've given us results for the face validity including discussions or discussing reasoning behind the disagreement.

So regarding face validity, 71 percent either agreed or strongly agreed that the scores obtained will provide an accurate reflection of quality and can be used to distinguish good and poor quality. They also performed some risk adjustments. There, however, was no conceptual rationale provided regarding the potential relationship between socialist factors and the outcome of interest, which here is the survival after heart attacks. They did note that clinicians responding to in-hospital cardiac arrest would not be a way of a patient's social economic risk and would, therefore, not be influenced by these considerations.

They also provided from c-statistics and R-squared numbers, all of – from their initial (inaudible) results from their 2011 to 2015 data – 2012. So all of the scores are very similar. They said this indicated that the initial risk-standardized survival rate model, it validates their initial risk-standardized survival rate model.

All right. There was some concern regarding missing data. They did say that missing – there was missing data in less than 0.1 percent of the patients in the registry. And then data on other patient variables has officially zero percent (of the) data. And also do not resuscitate status is not accounted for in the measure.

And for validity, we did have a split here. And I believe that was split mostly because some thought that this was actually a maintenance measure, it's not. So for a new measure, face validity is OK, is accepted by NQF.

So, we can start off there. Would anyone like to discuss any concerns they have regarding validity?

Jeffrey Geppert: This is Jeff. I think I can – well, in terms of like sort of cutting to the chase, I mean, I rated it low. I think given what you just said about the fact that it's a new measure and the concerns that were listed here, I – oh, I'd probably change mine to a moderate, so that would put us at the 3-5 threshold.

David Cella: Yes, that's helpful. This is Dave. Thank you. I couldn't understand why – where the concern was about missing data.

Jeffrey Geppert: Well, I think the only concern about missing data – I mean, and maybe this is an interpretation is my sense is that if the data is missing, this is not in the registry. So that would be a sort of a concern. I mean, it's not a concern to influence …

David Cella: Oh.

Jeffrey Geppert: … the rating but …

David Cella: Denominator in a sense not missing.

Jeffrey Geppert: If they data are missing, they just don't include it so …

David Cella: It's in cases.

Jeffrey Geppert: Yes.

Bijan Borah: And even then, this is Bijan. (Inaudible) they actually – I think they report it to be extremely low.

David Cella: Right.

Jeffrey Geppert: What's low if it's in the registry? I don't – I don't know if they reported it, like how many cases they don't even put in the registry. But it's not – it's not a big deal. I just …

David Cella: Yes.

Bijan Borah: OK, yes.

David Cella: This mean we're ready to vote?

Karen Johnson: Yes, I think so. This is Karen from NQF. Just the only thing that I would add to May's very nice description of this is face validity is definitely acceptable because it's a new measure, but you have to agree that the results are adequate so it's not just that they did it, but the results are good enough to satisfy you.

So, I'm assuming that's the case, but I did want to just be very explicit about that.

May Nacion: So just another quick confirmation, do you – do you want to discuss reliability or is everybody OK with their preliminary rating? Nobody seem to have any issues with their reliability testing.

OK. So we will take that as nobody has issues. You do not need to vote on reliability. Please just vote on validity.

David Cella: There is a vote – there is a vote requested on 2A reliability.

May Nacion: So, that is – excuse me. You do not need to cast a vote for reliability. You should be able to just cast your vote for validity and still submit the survey.

David Cella: Right. All right. Thank you.

May Nacion: Yes.

Female: OK.

Karen Johnson: If nobody else want to discuss anything about 3309, please cast your votes and then we can also move on to the next measure, which is 0964 (inaudible).

David Cella: Question …

Karen Johnson: Oh, yes.

David Cella: I may be doing it wrong. This is David Cella. But when I cast my vote for the first one, it kicks me out of the survey and thanks me. Do I just reenter or can I back there?

May Nacion: You can reenter the link by clicking that original hyperlink in the email.

David Cella: OK. All right. Thank you.

Female: Karen, did you want to start?

Karen Johnson: Yes. OK. Let's go ahead and start with 0964. So this is and maintenance measure that looks at therapy with aspirin, P2Y12 inhibitor – I don't know what that is – and a statin at discharge among patients who have PCI.

Now, this is a composite measure according to NQF' guidelines and it is an all-or-none composite. So there was a little bit of question about whether this really is a composite. It's not a traditional one that NQF does the all-or-none measures as composites.

The data source is a – the CathPCI registry. Level of analysis is facility. This is not a risk-adjusted measure. And it uses a combination. This is a little different in maybe some measures that you're used to seeing. It uses a combination of exclusions from the denominator as well as exceptions in the numerator.

So in terms of reliability, ratings were somewhat spread, three high, one moderate, one low. The low rating, we think, was due to disagreement about the appropriateness in testing methods and some concerns about the specs. And with that, what we would do since four of the five raters passed it, we will go through as either a high or a moderate rating unless somebody wants to pull the reliability discussion to discuss. So in terms of what they did for reliability, they did to score-level using the split sample methodology.

And just a note that because it is a composite measure, we do require score-level reliability should they hit our requirements there in terms of what they did. There was some concern with the specifications really that had to do with the way the measures work. The second component is a subset of the denominator of the other two, so that actually brought some concern that that measure – and this is sort of a validity question that the measure is really impacted not only because of whether or not people got the medications but just the frequency of PCI with or without stenting.

One other thing that we noticed is that the testing data were limited to patients ages 65 and older, but this limitation was not included at least that we saw in the specifications, so we probably just want to ask the developer to clarify that in the specifications. We could be harder on them and say your testing

doesn't match the specs and, therefore, you need to retest, but more than likely that I imagine you would be interested doing that.

They use the split sample method, and that also was a little bit of a concern, I think, with a couple of people. And they actually compared the Pearson correlation. That's a little bit different than what we usually see. And we did note that the split sample methodology to date has been accepted both by NQF in the past as well as methods panel in this past year. And it's something that we've talked about in some of our monthly calls. And I think we'll probably still talk about that a little bit more as to whether it's important what to actually require both of those or if we have a preference for one over the other. But again precedent is that that is acceptable and that if that is provided then signal-to-noise analysis wouldn't have to be. So that was reliability.

Again, let me pause to see if anybody wants to – well, let me say it this way. If you have no objections, we would just pass it. But if you have objections and you want to do some discussion and a vote then mention now.

Female: On reliability.

Karen Johnson: On reliability.

Female: (Inaudible).

Karen Johnson: Yes, we will still definitely talk about validity.

OK. Hearing none, really the trouble with validity so the votes ranged two high, one moderate, two low. And because of that split, we have to discuss and we have to vote.

So the empirical testing was at the score-level. And they basically did a construct validation comparing this measure to two 30-day mortality measures. And just to note, this is a composite measure and it is a maintenance measure, which means that the score-level testing this time around is required. So they did meet our requirements there.

We note that they also mentioned the face validity assessments that they did. What they did really doesn't meet what we are looking for face validity so the note there just is that don't really consider that when you're using the ratings. We really want you to consider the testing, the score-level testing that's provided.

Because this is and – well, I want to open it up for discussion of validity in just a minute, but just to kind of complete the overall look at the measure, it is a composite measure. Because it's a composite measure we have that extra criteria under the four composite measures. And again, the consensus there was that it would pass with two high, two moderate and just one low.

And I think the – what we put here in your discussion guide just to remind you are that when you're thinking about the composite questions for – to rate these, we want to know just the – do the component measures did the quality construct and add value? And then do the aggregation and weighting rules – are they, first of all, consistent with the composite construct? And do they achieve this objective of simplicity and to the extent possible?

So they computed hospital-level results for the three components and basically correlated those to the overall composite scores. And we've provided those correlation results in the discussion guide for you just to remind you.

So onto the things that we absolutely have to discuss about validity, again they did construct validation correlating their measures with two mortality measures. One was the STEMI shock mortality measure. The other was the no-shock mortality measure.

The correlations were in the expected direction that the developers hypothesized, which was better provision of discharge medications, they thought, would be associated with lower mortality. The difficulty is that the correlations were slightly low. The developer did provide a couple of ideas about why those correlations were a little bit low.

There was also, in terms of threats to validity, there was some concern about whether this measure would actually be able to meaningfully differentiate between providers. And we've provided here the mean, median and 25th

percentile. At the 25th percentile, the composite rate is about almost 92 percent. So the question there really is can it be used to differentiate providers.

And I think I'll stop there. So I think most of your discussions will hinge around that low correlation. And is it too low to fit you? And then the question about meaningful differences, the measures not topped out, but there might be not as much as room as you might like or this might be fine for you. And I'll just open it up for discussion.

Matt Austin: Yes, so this is Matt Austin. I was one of the folks who voted low on validity, specifically, because of the concerns with the low correlations. It didn't seem like there was really a strong relationship between this measure and the two measures that they proposed. I mean, there is a slight correlation and I guess I would welcome thoughts from my fellow panel members on why they think maybe that would be OK. I mean, I can be convinced, but just, I guess, I'd like some discussion about that and education maybe.

Lacy Fabian: This is Lacy. I have the same concern as Matt (inaudible).

Jeffrey Geppert: This is …

Bijan Borah: Hi, this is …

Jeffrey Geppert: I'm sorry, go ahead.

Bijan Borah: Go ahead, please.

Jeffrey Geppert: So this is Jeff. I think – so I was one of the high. And one of the reasons for that is I do think that there is just sort of the method and the result. And there are very few of these submissions that attempt to do what this developer did, which is to have a measure of the – of a quality construct and look at the association of that construct with a material outcome. I mean, a lot of people do like correlations with related measures and things, which I don't consider to be as compelling.

So I think I'm, in part, giving them sort of credit for doing validity the right way even though – and even though given the fact that they – as they say, there's a relatively low correlation. So I think I'm – I mean, just in terms of explanation, I'm sort of giving them – I'm giving them credit for their methodology and trying to signal basically that that type of approach is what everyone should be doing.

David Cella: Yes.

Bijan Borah: This is Bijan. David, go ahead. Sorry.

Male: Please go.

David Cella: Well, this might be a little bit of a shift of conversation. I kind of along the lines of what Jeff – so I think that was Jeff that said I wouldn't be as concerned about the local relations if the measure is actually potentially better than what it's being correlated with because it would represent an improvement in the system. I was actually more drawn to the issue that the 25th percentile is already at 92 percent. And with the times that I've seen NQF retire measures it's when performance get so good that it's not really a helpful quality measure and doesn't differentiate because almost everyone is doing it.

So is there something – I noticed this last question, any advice for the developers on how to improve the submission for this cycle or future cycle? Is there some way to build this measure out that would create more separation of providers?

I don't know the area clinically so I can't – I can't even speculate. I'm just wondering if – I mean, does – it seems like it's got potential in my mind, but maybe it would end up being sundowned within a couple of years of being implemented.

Karen Johnson: So this is Karen …

Male: (Inaudible).

Karen Johnson: … from NQF, and sorry to interrupt you. I just wanted to make sure that you guys are aware we do have another criterion called "opportunity for improvement." That criterion, which is considered by the standing committee, will be talking about that as well. And is it topped out, if there's still room for improvement that sort of thing. But it also comes up under a validity of the measure.

We didn't put the details in the discussion guide, but I think the developers did show some stratified results, which is another way to maybe think about that opportunity for improvement and how much difference there is. And perhaps one of the things that in the future submission mission should you guys go ahead and push it through, a future submission might be showing those statistics for the various subgroups if they didn't do it. And apologies, I don't – we didn't write them down here. So there might be a little bit of room in there that we're not seeing just by seeing these overarching numbers.

David Cella: Yes. Well, in a sense, what I'm hearing, Karen, is that maybe we don't worry about that because the parent committee will – they'll be considering opportunity for improvement, and so we don't need to – unless we consider that to be a major validity issue, we could – we could let it pass on that particular metric.

Karen Johnson: I think that's what I'm kind of hinting at, yes.

David Cella: OK, all right.

Karen Johnson: Yes.

David Cella: All right.

Karen Johnson: And I will kind of take it back to the composite thing, which you guys did say you've felt the composite was OK. One of the things that we hope that you looked at when you thought about the composite is is any one of those components kind of pushing things up. If there are too many components in there, and I don't remember what the numbers were split out, so may not be. But that question is actually considered under the composite criterion. But we can certainly, if you guys do decide to go ahead and pass the measure, we

would certainly note for the committee that the higher percentages or the higher rate was a concern even though you passed it for.

Matt Austin: So this is Matt. Jeff, I appreciate your comments about the testing method that the measure developer use. And I would agree this is exactly what we want others to be doing, so I completely agree with you there. It's just when it's not very far from zero, it feels like what does that really tell me about this measure. That's the only thing there.

And I agree, I'm not sure one would expect necessarily agree each relationship between a process measure and this outcome measure, so outcome measures that they chose. So a little bit of my hesitation was did they really choose the right measures to compare against?

Jeffrey Geppert: Yeah, that would – that would be good feedback as well.

David Cella: Yes.

Jeffrey Geppert: More proximate outcomes that they could use in their validation.

Karen Johnson: OK.

Matt Austin: Thank you.

Karen Johnson: Any other discussion along these lines or anything else that came up or you – do you feel still ready to vote on validity?

Bijan Borah: So I have a question, this is Bijan. So what is that – the norm in terms of the correlation coefficient for the components with the composite? So I – you know, the – for P2Y12 it is 0.89. I guess, the question is, I mean, what is the threshold?

Karen Johnson: Yes. Well, that's a good question. NQF doesn't have any thresholds. And apologies, I haven't looked at this measure in a couple of days. I think probably what this correlation is telling you is that the statin component is really, really highly closely correlated with the results of the – of the outcome or the composite as a whole.

And often when you have correlations that aren't as high, the argument is that they're related so they kind of belong there, but they're not completely driving things. But I wouldn't go so far as to say and maybe somebody else could help us on the call is that 0.95 telling us that the statin component is needed or it's kind of superfluous. That I don't know the answer to, and Jeff might know.

Jeff, you do a lot of composites, right?

Jeffrey Geppert: Yes, I've – I have – in general I find those types of kind of item correlations difficult to interpret in that – in that way because on the one – because I'm not sure it really answers the right question. So to my mind, there's sort of two sort of fundamental questions that a composite sort of needs to address. I mean, the presumption is that relative to this, the individual measures, the composite sort of adds new information that the individual measures don't.

So that's sort of a concept of kind of competing importance. You could have one provider that was high on one and another provider that was low on the other, and how are you supposed to make a decision in that context. And that composite could potentially inform that type of decision-making, make a more rational decision than someone could just on their own.

And then the other sort of theoretical sort of rationale for composite is uncertain component – uncertain importance when you don't know at the time you're making a decision which of those components measures is the most important. And so it's based on kind of the probability that it will become important in the – in the context. And I have – I have a hard time sort of taking those sort of cell-level correlations and associating them with either of those rationales.

This is a – this is in all or – this is an all-or-none, right, sort of composite?

Karen Johnson: Right, yes.

Jeffrey Geppert: So I just tend to – I tend to – from an evaluation perspective, it just makes more sense to me to treat the – treat it as if it were a process measure with some (inaudible) and logic (and) the numerator …

Female: Yes.

Jeffrey Geppert: … because that – to me then that just makes sense. This is a process measure and you're evaluated like you would in any other process measure.

Karen Johnson: It was a …

Jeffrey Geppert: Right.

Karen Johnson: … a point of great discussion back when we decided that all-or-nine would be considered as composite so.

Jeffrey Geppert: Yes. I think one thing I suggested that maybe would be more useful than these kind of correlation tables is just to get me like what's the especially if they have these kind of outcomes data, what's the outcome in each of the nine cells.

Male: Yes.

Karen Johnson: The actual frequencies if you aggregated to the facility level?

Jeffrey Geppert: Yes.

Karen Johnson: Yes, OK. OK. I think we hit lots of things on this measure, and we've a lot about validity. I'm going to give you about two seconds to chime in. If not, I'm going to ask you to vote on validity.

OK. So clearing throat doesn't count as chiming in, so go ahead and vote on validity and we'll move to the next measure.

Female: Right.

May Nacion: Well, thank you for that. So (inaudible) and I'll be talking about Measure 2936, Admissions and Emergency Department Visits for Patients Receiving Outpatient Chemotherapy. I think the biggest confusion about this measure was if it's a new or maintenance measure. This is a new measure. However, it has come through the endorsement process before and it's not being seen by

the Methods Panel, but it was reviewed by the Cancer Standing Committee in 2006. However, at that time, it did not pass on reliability and thus was not endorsed. And so when we're reviewing this measure we should really review it as a new measure.

So this measure estimates hospital-level risk-adjusted rates of inpatient admissions or E.D. visits for cancer patients 18 or older for at least one of 10 conditions within 30 days of the hospital-based outpatient chemotherapy treatment. So the rates of admission and E.D. visits are calculated and reported separately.

I do want to bring up there were some comments about why are these two – one measure, maybe it should be two measures. Maybe they should be evaluated separately. This is something that came up during the standing committee discussion the last time it was reviewed. And both the developer and the standing committee agreed with a greater majority that it was better to just leave it as one measure. And so that's why it's now come back for as one measure.

This is an outcome measure based on claims and enrolment data at the facility level and it is risk-adjusted. So in terms of our reliability, this measure did pass with a moderate rating. We had one high and four moderate. There were some concerns about low reliability in non-cancer hospital. There were score reliability done in two autonomous, signal-to-noise and a split-level ICC. The results there are listed for you.

So I'll pause real quick since it was a pretty big majority that agree that the reliability testing was appropriate now. I'll ask does anyone want to pull it for discussion. And you can always pull it later if I'm going too quickly, but I am going to assume that we're good to go with reliability and we can focus on validity.

So in terms of validity, we did have a split decision, two moderate, two low, one insufficient. However, much like our earlier discussion, I think this maybe partly because there was some confusion about if this is a new or

maintenance measure. For new measures, we do accept face validity as the method for validity and empirical testing is not required.

In terms of the face validity provided for validity, there were some concerns about not actually – not all of the things are listed on the face validity accounting of face validity, and that's true. With NQF requirements, only the mention of the 28 expert workgroup meet our requirements. All the other attempts and groups that they had reviewed the measures don't really count because they're more about the process of developing the measure and not actually evaluating if they think it's appropriate. However, the 28 extra workgroup does meet our requirements.

Also with that, there was some mention about of the eight respondents. There were a couple that were involved in the development. However, NQF does not have a requirement that they use it completely independent. So that's actually fine for them to be using those people or part of the development.

Other than the actual discussion about method of testing, there were some concerns about risk adjustment brought up about is concurrent radiology or risk factor present the start of care and is there any potential for complications from radiology with complications on chemotherapy. The developer does state that concurrent radiotherapy is defined as having a radiotherapy procedure present on the same claim as the first index chemotherapy case or on a separate claim within 14 days prior to the first index chemotherapy case.

With that, the only note I will add is that the inclusion or lack of specific risk factors should not be a reason to reject the measure, although concerns can be raised to go to the standing committee who can then decide that issue go down for that factor. But at the Methods Panel level, it wouldn't be appropriate to drop the measure for that.

And then again, there some mentioned about the two ways being combined into one measure. And again the Methods Panel can state that they think that's not properly done, but if they feel that they provide extensive analysis and discussion about the consideration of the social risk factors and the standing committee has data that they feel is appropriate for the measures to

be won, it's really up to Methods Panel to decide how they want to move forward that information.

And lastly, there were some concerns about meaningful differences, and we have listed the number here for you. So with that, I will open it up for discussion to see what are some thoughts on validity for this measure.

(Inaudible)

Male: I'm sorry, go ahead. OK.

Matt Austin: All I was going to say was I think I was – I had raised a concern about that some or all of their expert panel that they had reviewed the measure for face validity had been involved with the development of the measure. And I can appreciate that NQF doesn't have specific rules around that since I'm happy to accommodate that. I think that might be something that NQF might want to think about for the future. It just feels like there is a strong conflict there if you're part of the team developing the measure. I'm not sure you can really independently assess whether or not it's a useful measure for quality and safety. So that's just my two cents.

Karen Johnson: So this is Karen from NQF. And just FYI, a few years ago we did try to add that to our guidance and criteria that the TEP evaluating face validity would have to be separate. And we got a lot of pushback from measure developers, I think, particularly in some cases where the clinical expertise is limited to a fairly small number of people, so their argument was that, A, we got the greatest people to help us develop it, but there's not enough people that go out and find independent groups to evaluate it. So that was the argument.

And again, I think that was probably back in 2013 or so. It's been a little while so we could certainly reconsider it and see if that – if it's time to be a little bit more stringent on it, but we have tried in the past.

David Cella: So this is Dave Cella. Karen, so to the – to the group, to the – to the team, I was co-chair of the Cancer Committee that reviewed this a couple of years ago, so that's a disclosure. And my question to you, Karen, is that – is it OK

for me to mention that some aspects of that discussion in this context or would you rather keep that separate?

Karen Johnson: No, I think you are perfectly free to other than, yes.

David Cella: OK. So on face validity, and correct me if my recall is wrong, Karen, but my recall is that the committee was – had no problem with face validity. It's – I think it's a pretty common shared view across oncology that keeping people out of the emergency room and out of the hospital is a good thing.

There was a lot of discussion about why they're combined and not separated. And actually that expanded beyond that because there are so many different models of care now that is possible not only to avoid hospitalization by treating an emergency room or 23-hour stays or to avoid emergency room care in certain places by having urgent care – urgent care, which may not reflect the – your patient pool being any healthier, you're just coding them as urgent care. So those are some of the concerns.

I don't think the committee was concerned about combining at the end – at the end of the day combining E.D. and hospitalization, but actually that there are still other models of care that would allow you to bypass coding something as E.D. or hospitalization in patients who otherwise are similarly sick. So did they address that at all here?

Karen Johnson: I don't recall that being addressed when I look at these, Dave. I don't know if other panelists know it. And it – I will kind of an NQF caveat, too, sometimes it gets tricky. Our numbering system isn't as pristine as we would like. And over the years, we actually have allowed these kinds of what we kind of now internally call multi-rate measures to go through. You're used to them very much when you see a CAHPS measure or something like that where there's 11 different measures all under one NQF (I.P.).

In this case, they really are and we do consider them two separate measures so we expect testing, et cetera specifications, all those things to be laid out and separated out, which I think they've done in this case. So it's under one NQF number, but it really is two measures.

David Cella: Right, yes. Thank you for that. That's important. So I guess I just wanted to comment that so in terms of face validity, there is the independent record of the Cancer Committee, the NQF committee that didn't have any conflict on it as endorsing the face validity.

Matt Austin: Yes, thank you. And this is Matt again. I hear that. I guess I – just from raising the concern about conflicts and I would just push back and say face validity is one option that they're given on how they would demonstrate validity, and it's obviously, in some way, sort of the easiest but there are other options that are available as well.

David Cella: Yes.

Matt Austin: But I'm – it'd be possible to hear that NQF doesn't ask specific criteria around that methods conversation has been brought up before and people have been comfortable with having different viewpoints. That's helpful. Thank you.

Karen Johnson: So this is Karen again, just a couple of things. If you guys do decide to pass the measure from your end, and it actually goes through on the standing committee end, when they bring it back in three years for re-endorsement, at that time, they would be expected to provide empirical validity. That's one thing that we have changed actually in the last couple of years is that face validity we will allow for new measures but at the time of maintenance we really would like to see empirical testing.

And, Dave, to go back to your comment, it sounds like that you would suggest that developers consider other forms of care that might also be added to, at some point, this measure. Is that – is that a fair statement?

David Cella: Yes, but I don't – I don't take it as – in itself a knock against validity. I'm not arguing validity …

Karen Johnson: Right.

David Cella: … but I do – I do – that did come up and what's the concern expressed. I'm sure the committee will be wrestling with that again.

Karen Johnson: Yes.

Miranda Kuwahara: OK. Were there any other concerns or topics that you want to bring about validity? Otherwise, we can go ahead and vote? OK.

So please go ahead and put your vote only for validity for 2936. And then we'll move on to our next measure, which I think is Dr. Karen.

Mike Soto: Just to let you know, this is Mike Soto. I'm finally joining. I'm sorry I was late. I won't vote on anything up until now.

Karen Johnson: Thank you, Mike. That's great. Glad you could make it.

OK. So we are now ready to go to Measure 3478, Surgical Treatment Complications for Localized Prostate Cancer. This is a new measure and it's basically an interesting measure. It is using claims to look at kind of pre-surgery versus post-surgery, urinary incontinence and erectile and/or erectile dysfunction amongst patients who have localized prostate cancer surgery.

And they build this to where the outcomes are rescaled to a zero to 100 scale. We think that this measure is meant to be stratified by – let me try this word prostatectomy type. I think that's surgery type, so we would expect some results split out by the stratification there. And the measure specifications are guessing that the measure be limited to facilities who have at least 10 patients attributed.

So, the level of analysis is facility. This measure is not risk-adjusted. And in terms of the weightings for reliability and validity, there is actually splits on both reliability and validity. Reliability was tilting a little more towards the low side. The concerns had to do with specifications and the testing methodology. And for validity, there was a split tilting a little bit more towards the pass side. There – in terms of what they did, they did a little bit of data element validation and a face validity assessment, which again is – because this is a new measure that would conform to our minimum requirement, but there was also some concern about exclusions and risk assessment.

So with that, since we definitely have to discuss both, we'll talk about reliability first. In terms of specifications, there is some uncertainty about the truncation and rescaling and how that was actually done. So that wasn't kind of brought out in the submission.

And the other question that came up and it actually comes up, I think, both under specifications and reliability – I mean, I'm sorry, validity is just a curiosity factor about how many hospitals aren't actually eligible for the measure since they have a volume – a minimum volume provided. And that was information that they didn't tell us.

In terms of this testing that they did, this was a – the split sample methodology again comparison via a Pearson correlation, which was it came out to 0.65. Again just a reminder, I think everybody said this today that the split sample methodology has been accepted in the past on its own by NQF and the Methods Panel. But the developer this time didn't really tell us much in terms of how they split the sample. We know it was a random split, but the details were a little bit lacking there.

And again, it seems like this is meant to be reported separately by open versus not open surgery. A question for you, as a Methods Panel, would be do we need to see reliability results without by surgery type or having them all together in one – in one analysis? Is that OK? So, what I'd like to do is stop there and let you guys talk about reliability first before we proceed on to validity.

Jeffrey Geppert: Just a quick comment on the split sample thing, so I think that's just obviously an issue that we need to address because I'm definitely in the camp that it's not telling us what we – what we want to know with respect to reliability, but I understand what you're saying about the fact that it's been accepted in the past and is an acceptable method.

Karen Johnson: Yes, and this is Karen. We will definitely be talking about that again in the next few months. So …

Mike Soto: Hi, this is Mike. I've got another issue. My understanding is that when they had 10 patients or fewer then not that surprisingly they have reliability

problems. Is this intended only for bigger hospitals in that or would it be used for small hospitals, too, which there seem to be many in the database?

Karen Johnson: So the idea is that a hospital that had at least 10 patients that fit their denominated criteria, this measure could be used for those hospitals. But if they only had nine or eight or something along those lines, then it should not be used by those hospitals.

Mike Soto: Yes, but I – even for 10, it seemed to be pretty much of an issue. They – I'm looking at my notes here. They say for at least 40 patients they have a correlation of 0.85, but then it's only 0.65 for hospitals with at least 10 patients, and there seem to be a lot of those.

Karen Johnson: So what did other panelists think? Correlation is getting pretty high with 40 or more, somewhat low with 10 or more. I will tell you yesterday – was it yesterday? I'm already getting mixed up on when we had our Subgroup 2 call, the discussion did come up a little bit about the split sample methodology versus the signal-to-noise methodology. And I think the consensus from the Methods Panel or at least there wasn't really pushback from it, the consensus was that that number that would come from a split sample should be regarded as a conservative number, almost maybe a lower bound. Again, that was a statement made by the developer for one of the measures that was considered by Subgroup 2, but the group did talk about it and again seemed to kind of accept that thinking.

So I don't know if that will help you in your thinking today or not, but I did want to share that.

Jeffrey Geppert: Yes, I don't know about that.

Karen Johnson: I've got it on my list where additional monthly call, so it's definitely something that we want to tackle.

Jeffrey Geppert: Can you sort of repeat it again, the issue you raised about the fact that there's sort of open and closed and – I guess, I didn't completely catch the other reporting two different rates.

Karen Johnson: Yes, I mean, this is not – this is Karen again from NQF. The way that they stated in their submission, it sounds like that they would report this out for open surgery versus closed surgery, so even though they didn't say, "Hey this is two separate measures under one NQF ID," in reality, if they are going to compute them separately and report them separately, then that's what you have.

So my question for you is, knowing that they would plan to do that, do you feel like we would need to see this reliability numbers for the open surgery sources and the close surgeries?

Jeffrey Geppert: I wouldn't say yes if that's how they are going to – how it's going to be reported.

Bijan Borah: Yes, this is Bijan. I agree. If they intend to do that, if they encode it by open versus a robotic or a non-open, then probably, that's how it should be encoded as well.

Matt Austin: And this is Matt. I'd agree. I didn't catch that in the stratification section. I sort of glanced over that.

Karen Johnson: Now, it would be really helpful if – I don't know if any of you have your measure information sheet opened, if somebody could definitely confirm if that was the intent of the measure and I'll try to open it as well. We don't want to …

Matt Austin: Yes, it's Section F on Page 5. It says, "Each hospital's performance score should be reported with this measurement and serve as the basis for national comparison and accountability. However each hospital's performance should be reported stratified by pressing the technique type, open versus not, to add meaning for consumers and for hospital quality improvement.

Karen Johnson: So it does sound like from that statement that in terms of the accountability and public reporting, they would plan on splitting them out. For internal (PY), I think we would be as concerned.

Matt Austin: Yes, I mean, they are slightly different populations and as I understand it, may reflect some different degree of disease or have differences, so one might actually see differences in urinary incontinence and erectile dysfunction – these are types if I understand that, so …

Karen Johnson: One of the things about doing is if you're pretty happy with those reliability results as a whole, but your concerned about the stratification, you could forward that audit to the standing committee to decide. It's kind of up to you where you would land on that.

Matt Austin: So, this is Matt, let me pose a question to the group, so I thought that this is moderate. They did use the split sample half – split sample approach which NQF has deemed as OK, and there was a correlation of 0.65, what am I missing or why would others not see that as moderate?

Bijan Borah: This is Bijan, I voted moderate, too and because of that, that's like – I think as far as I understand, 0.65 is considered sort of fair. So, again, I mean I always get confused – not only confused, I mean, I have no idea in terms of what is the (trends) or what is a number that it ties it to consider good in terms of the correlate and coefficient.

Jeffrey Geppert: This is Jeff. I am intending to go from a low to a moderate for that reason.

Matt Austin: OK.

Bijan Borah: But I know, I actually agree – the fact that – this is again, I didn't get it earlier, so if they are intending to report it separately for open versus robotic, I think by definition, we all know that open surgery probably would have higher complications and I would really like to see the numbers certified by open versus non open.

Mike Soto: Also, if they reported two different rates, each one presumably will have less reliability than the number together.

Bijan Borah: Yes.

Jeffrey Geppert: And fewer hospitals would meet that 10 threshold, so …

Female: (Inaudible) especially good either …

Mike Soto: But what I'm concerned about is that, a hospital that may have had 25 will now have, 12 and 13, but those will be less reliable.

Matt Austin: The panel does want the developer to split it up between open and not open, does that design influence our reliability vote or is that just a note back to them or …

Karen Johnson: I think it's a judgment call on your side if you push it through as moderate than that would be saying you're OK for now with kind of the grouped reliability and the results that you saw, but it sounds like, at minimum you would want a very strong statement saying when it comes back, or perhaps even more time close in time, we actually could say, "Hey, this is something that we'd like to see maybe in the year," and you could put that forward as a very strong desire on your part.

So, no hard and fast rule on this one. It kind of depends on are you willing to see it endorsed potentially with this number or is that a complete no go for you?

Bijan Borah: This is Bijan. I would go for the letter of outcomes current – the one where you sort of let them know that we would like to see the numbers certified by open versus non-open in year's time.

Karen Johnson: OK.

Jeffrey Geppert: I agree with that.

Karen Johnson: That sounds like a reasonable ask from the rest of you, I mean …

May Nacion: I do want to comment that while the recommendation can be made and the standing committee can hold that recommendation, there is no hold – they don't necessarily have to come back with that information. They can do during their ad hoc and then see it from there.

Karen Johnson: We've had precedent. I think, we've asked that they need to bring things back, so I think this one might be one that we could expect that they could do that for us. Or if you're willing to wait three years, I mean, that's OK, too.

OK, yes, that's – it's a tough question. So go ahead and vote however you would to vote and we will capture these comments in our summary.

Mike Soto: This is Mike. Since I came in late, I don't know how to vote. Can someone …

Karen Johnson: Yes. (Sure). Miranda has got a frog in her throat today, so she's going to hand it off to May.

May Nacion: So we sent you an e-mail yesterday with the meeting materials and at the very bottom of the e-mail is in hyperlink to the SurveyMonkey link, so if you open that up and then click on this measure which is 34-78, you can vote from there.

Mike Soto: OK.

May Nacion: (Inaudible) (if there's any) problem.

Mike Soto: And we're just voting reliability, right? Because I don't think we've talked about validity yet?

May Nacion: Correct.

Karen Johnson: Correct.

Karen Johnson: OK, in terms of validity, there were some general concerns. We've already talked about not really knowing how many hospitals are going to be excluded from the measure. But just having that would be interesting information. There was some question about the score being a scale from zero to a hundred, and the comment that a hospital can seem to get worse simple because their hospitals get better. There's a question about how facilities would interpret the score when the rescaling is done at the patient level?

If that's something we really want to discuss, the person who made that comment might want to explain that a little bit more clearly so that everybody gets it and then, there was some additional desire for additional analysis or explanation around exclusion.

In terms of the testing, I think it was unclear whether the face the validity assessment actually met our requirement and the – just to reiterate what our requirements are, we are actually looking for a systematic assessment that is transparent by the identified expert and we want it to explicitly address whether the performance score is from the measure can be used to distinguish between good and poor quality and then we also want discussion about what was the rationale if the actual was in the negative.

So, for this one, it seems unclear about the size and composition of the TEP. The questions that they asked weren't exactly what we say in our guidance materials. They did ask statements that they made where the performance measure succeeds in measuring what it was intended to measure and the scores reflect information regarding the quality of prostate cancer surgery, that latter one may be especially very much close enough to what we are asking for that you're fine with it.

Results were – for the first question, a four out of five, sorry a 4.5, seven out of eight agreed or strongly agreed with none disagreeing or strongly disagreeing. For the second question, the average was a little lower, five out of eight strongly agreed or agreed and again, none disagreed or strongly disagreed.

So when you're thinking about the face validity assessment, are the questions close enough to what we're asking and are the results reasonable, adequate enough to you? They did some data element testing. They actually looked at the cohort definition and what they did is compared it to the (CR) database and kind of verified that they were able to find the right cohort for the denominator.

In doing this comparison with the (CR) database, which they considered the gold standard. They didn’t compare other critical – other data elements and

just a reminder that when data element validation there, then NQF would like to see all of the critical data elements with that, so in this case, they didn't look at all of them, but again, this is a new measure, so not looking at the critical data element, all of them would have been a fatal flaw potentially for a non-new measure but with a new measure, the fact that they did and the face validity assessment could push them over the passing edge.

There was also – I think a little bit of concern about the lack of risk adjustment. That wasn't gut out very heavily in the comment. But the developers did spend some time talking about when it's just not to adjust.

So, with that, let me stop to see if anybody first has any questions on things that I mentioned and then we'll just talk about it.

David Cella: Well, this Dave. I am concern that erectile dysfunction and incontinence are not – they are graded on a level of severity and this conversion from zero to a hundred, I think is – it's not clear to me that it is valid just as at the element level.

And I don't know, risk adjustment seems pretty important here, especially if some of these men are also going to be getting radiation or hormone therapy.

Matt Austin: This is Matt. My understanding of the risk adjustment and I could be incorrect in this was that they actually did look at it both ways and then did a correlation between the two and found that to be 0.95.

So my understanding once they were sort of coming to the conclusion that the risk adjustment didn't really change performance results and they …

David Cella: Right, I just have a hard time buying it, I guess.

Matt Austin: Right, and maybe – and I didn't necessarily assess their risk adjustment model and whether it included all appropriate and relative risk factors, so if it was an undeveloped model, then perhaps, that could be a reason.

Karen Johnson: So I think in terms – just to give you a little bit of guidance in terms of measures that maybe you have to at least consider risk adjustment. What we

ask developers to do is if they decide they are not going to, they need to make the case as to why they chose not to.

So in terms of your decision there, you're basically – we're asking, do you accept their justification as Matt said, their correlation analysis. Is that enough to convince you that risk adjustment probably wasn't that necessary. Is that how you would approach that number?

In terms of this, you wrote to 100 scale on – I have to say, I didn't quite understand what they were doing, so I don't know, Dave, if you could even – can you explain? Do you …

David Cella: I can't – I am just very skeptical. I'd rather have it be much more explicit about what they're getting from the patient in terms of their report. I think that if these charts, if they're documented, then they have some conversion, I do not understand it either.

Karen Johnson: Is it because different people have different numbers of claims and can make it fair to compare. I mean, I don't know, I am asking. I don't understand.

Jeffrey Geppert: One of my issues was that they sort of interpret this as the mean the mean difference in days. I mean, that's sort of their interpretation of it, although the score essentially eliminates that interpretation.

And there was scaling that is done at the patient level, so I don't know how they can refer to it as a mean. It's not actually a mean of the difference in days, it's a mean of the score, which is rescaled in some unknown way. And this wasn't clear how they were even calculating claim days.

David Cella: I mean, it's a really good idea performance measure, I just can't see how it's going to generate any confidence that we're really getting at, now the proportion of people that have clinically significant incontinence or erectile function. The numbers will be higher than what is reported, and it's hard to know how much higher.

Karen Johnson: So this one is cheeky too, so it sounds like at minimum, whether it goes through or doesn't go through from you guys having some additional clarity at

some point on the scaling methodology and how that works is something that you would like to see.

It also sounds like there is a little bit of hesitation just in terms of the clinical side of things that if you guys did go ahead and pass it through, we would want to ask these clinical – make sure that the standing committee hits those clinical questions.

Jeffrey Geppert: Yes, I think that – I was going to make a point along those lines, too. To some degree, we are making – we are thinking about this from the clinical term as opposed to a statistical approach and I'm not sure – I know that I'm not capable of thinking about that without having substantive knowledge.

Karen Johnson: And that's extremely fair. We don't – we want to make sure that the standing committee is able to put some of these questions in context, so it might be that in terms of what you're looking for, you'd really want to be looking at just the face validity assessment that they did – is that adequate for what you'd be looking for? The cohort identification seem to work pretty well and so regardless of the fact that they didn't look at all the clinical variables, the cohort definition did work, which was a major question that they asked and they did provide at least some rationale as to why they did not risk adjust and that kind of leaves the clinical question about should – are there other things that they should have considered would be something that maybe the standing committee could weigh in on.

Is there anything else you guys want to talk about on this measure? It's a tough one actually.

Jeffrey Geppert: I do think one thing that would be helpful is if – they could just even provide like a table and so – I mean, I think they start off with a fairly simple concept and then they really make it complicated, and so, don't do that or if are going to do it, just make it easier to map it.

Like, I don't know how their panel could answer the question without knowing – does a score from 90 to 92, what does that mean in terms of days? What's the mapping between their reasonable construct and the score they've created.

Karen Johnson: OK, we're writing this down.

David Cella: Yes, I second that. This is Dave.

Matt Austin: And this is Matt. See, one thing – and I don't know if this is the right place to bring this up, but I do have some concerns with them excluding patients who die within the year of prostatectomy. That's a little bit of a potential survival or survivor bias that's introduced.

Karen Johnson: OK, it sounds like the well has dried up on discussion on this one, so a lot to consider. I'm going to go ahead and ask you to cast your votes on validity for this measure. And we'll give you a minute or two to do that.

David Cella: How are we doing? We've got a half hour left. We've done five with three to go is that right? Or have we done four?

Karen Johnson: We've done four. So that actually is about what I expected, so in terms of what I was expecting, we're a little bit ahead of time. I kind of doubt we'll get through all four of them, and that's OK, that's why we set up a second call.

So let' go ahead to 25-61 and for this one, Ashlie is going to walk through it. So Ashlie, are you there and maybe to take on 25-62?

Ashlie Wilbon: Yes, I am ready. And Karen, I have to say I am going to be optimistic about this. There are four measures left, two – if you want to look at it as kind of two sets of two, the first two measures are very similar with similar issues and I think, maybe if we get through the first one with some good discussion, we can talk about how on the second one maybe similar or different and we will need you to vote distinctively on each measure, but I think we may have a little bit of time savings given that we have similarities in these two (inaudible).

So I am still hopeful we can do it. We'll do our best. So the first measure that we're going to look at is 25-61, which is (SCS) aortic valve replacement composite score. It is a maintenance measure and their first line of the description, they describe it as two domains consisting of six measures, but I

think that the easiest way to think about it is two domains, one of which is a 30-day mortality and the other major morbidity, which that back component has five – within it, five adverse outcome that they're looking for that are kind of the absence of or the presence of within each domain, which within each of those five domains in that element.

So it is calculated as a weighted average of the two domain estimates where the weights are inversely proportional to the standard deviation of the domain specific scores across hospitals. There were – never mind, I'm going to stick to where – I was going to veer off here, but we'll stay here.

The measured timeframe is three years and facilities and groups are excluded from the measure if they have less than 10 AVR procedures in the patient population, so this measure relies on the registry data from the (SCS) adult cardiac surgery database. It specified the level of analysis at the clinician group level and the facility level.

For reliability, the panel passes measure potentially – basically with a moderate rating. There were five moderates and I'll go over some of the things within this reliability element, but we don't necessarily have to discuss this in detail if everyone is OK with their ratings, unless someone wants to pull it.

So in terms of reliability, there were – extensive specifications were included in the submission date. They did provide a link to an external PDF document that describe their specifications, but it's not actually in the meta-information form, so hopefully, you guys were able to locate that. I will work with them to make sure that the specifications can actually get – and the measure information form.

In terms of reliability testing, they did core level testing using signal to noise ratio and we listed the results here. They have divided them up among participants with 50 or more operations and a hundred or more operations, though posterior mean of reliability across all of them was 0.49 and the posterior median lower and upper boundaries were 95 percent credible intervals was 0.490 with a range of 0.44 and 0.54. For participants with 50 or

more operations, the reliability score was 0.59 and for participants with a hundred or more operations, the reliability score of 0.69.

A note for the panel that NQF requires that for each level of analysis that they specify it in the measure. They should provide corresponding testing and that – based on how they have described their testing, it's not exactly clear which of their results correspond to each, so that's something that we will need to follow up with them on – and in terms of clarifying their submission to make sure that either that they need to better label their testing or divide it up or provide additional testing that corresponds to whatever is missing.

So I am going to pause there and see if anyone has any additional comments or issues about reliability that would like to be discussed. Again, it did pass with five moderates, but I just wanted to open that up for discussion.

David Cella: So are you saying that even if we just go ahead and say that five moderates, this passes, kind of gets – you're still going to require some additional testing?

Ashlie Wilbon: Potentially, yes. If it passes on validity, and the measure gets through the method panel, we would then – we would file them with the input from the panel anyway, but yes, there would be additional information that would – or clarification that will be needed from them in order for this to go forward to the committee.

David Cella: OK.

Ashlie Wilbon: OK. Hearing nothing on reliability, let's move on to validity. I did also want to make a point just of consistency because this measure in very similar to the measure that we will be discussing next, so just kind of keep that in mind. We vote here potentially and then vote on the next measure that if there are reasons that your votes are different between the two measures given the nuances between them and that show your differences, that's fine, but where there should be consistencies, I just wanted to point that out that we should try to be consistent as possible across the two measures.

Especially given that the issues with validity that were identified were essentially the same for the two measures, so with that said, I'll go ahead and

move into validity. There where – the votes were for three high and two low and again, this puts a setback kind of three out of five split which is right on the cusp of what we consider consensus.

The measure score validity was assessed using face validity of the composite, content validity of the components and predictive validity to show stability over time of the composite. Most of the concerns from the panel members were around the appropriateness of the predictive validity method, and we will come back to that in just a second, but also there were some concerns around the approach to determine the inclusion of (STS) factors in the risk adjustment model. This was primarily around the concern of them including race, which they described using it as a clinical factor, which has some relationship with genetics and the expression of the disease in some racial groups more than others.

Also concerns regarding meaningful differences and we'll come back to that as well. And then again, this note about the level of analysis, which we've already discussed. So let's pause there and I'll just highlight a couple – again, a couple of issues with validity that came up and then open it up for discussion.

With the face validity assessment that they provided, we've talked about just a little bit already, but their description of face validity doesn't quite meet NQF requirements on what is needed in terms of providing some sort of results or other description specifics about how the experts rated or determined that the measure score of what they were performing distinguished good from poor quality.

That said, they did provide other types of validity testing which should be the focus, so they describe a predictive validity assessment which examine the stability of Star ratings over a three-year period and they have divided the performance of the participant in their sample, classified them as one, two and three-star participant based on their scores. The greatest stability was found among those with two-star ratings. This was also the group that had the most participants in that group, and so there was some concern over whether or not

that would actually be expected given that was the largest group or aggregation of performance within that two-star rating group.

With the predictive validity assessment, they compared the components of the composites to the – I'm sorry, they compared the components of the deposits of morbidity and mortality rate within each of the Star-rated groups.

So I will pause there and open it up for discussion on validity.

Matt Austin: This is Matt. (Inaudible) first, I was going to raise the concern about their predictive validity and my concern is that, with 90 percent of hospitals being classified as two stars, one shouldn't be greatly surprised that in the next round, a very, very high number of them are still classified as two stars.

And so, I don't know what the – what I am proposing as an alternative, but what they've done feels a little short of what I would expect because I feel like it's sort of the expected results given how they classify hospitals. But I welcome pushback and feedback on that thought.

David Cella: Does it taint the measure. I mean, that totally makes sense, man, I mean, you're going to – if you're 95 percent base rate, you're going to have high stability, but does that really paint the validity of a measure in your opinion?

Matt Austin: It doesn't change the validity of the measure, but from that, I am not sure that they demonstrated sort of the validity that they were hoping to demonstrate.

Mike Soto: This is Mike, I guess, I thought that the real validity test was the graph that shows adjusted morbidity and mortality within the one, two and three-star groups. This is just above 2B-1.4 on Page – I don't know what page it is.

I think that's the relevant comparison, isn't it? That the morbidity and the mortality are so different between the one, two, and three-star groups?

Matt Austin: Yes, let me look at that right now. I may have …

Jeffrey Geppert: Those aren't independent though, are they? I mean, aren't they just showing that they constructed the composite by …

Mike Soto: No, I got the impression that the stars were constructed on other grounds than the morbidity, than the measure. Maybe that's not – it's not totally clear about that, but that's the impression.

Jeffrey Geppert: The sentences that they – the composite scores have higher performance on each individual domain, so what they're doing is they're showing that when the composite is higher, the component measures are higher, which again doesn't really seem to be – but validity and I mean, a measure is persistent. It could be persistently biased. I mean, it doesn't basically say anything about validity.

Matt Austin: Jeff, this Matt, are you saying that a stronger comparison would be that compared to their measure to an independent measure, as opposed to measures that actually comprise their measure.

Jeffrey Geppert: Right, yes.

Mike Soto: This is Mike again. I did look at it and I think that I was wrong and that you guys are right that the stars are based on the measure, so that your concern is valid.

Female: Yes, the same.

Jeffrey Geppert: I mean, eventually …

Ashlie Wilbon: So go ahead, sorry.

Jeffrey Geppert: Sorry. I just said, I'd like to see us sort of progress towards kind of a – it's almost like sort of a mature level with these kind of validity studies. I mean, to me the lowest level of kind validity study is to say, there's some sort of implicit quality construct. We can't measure it, we just sort of suggest that it's there and that could be because it's consistent over time or there is some variability across measured entities and outcomes. So we're going to sort of assume it's there, it's kind of an implicit argument.

The second kind of – so that's to me – that's like the lowest of the low. It's something, but it's low. Our more moderate level will be to say, we still can't

measure the implicit construct, but we can create the construct and we can associate it with some material outcome and that's kind of what some of the ones that you did earlier, but you can imagine, you have a composite, you could actually – you could actually construct the composite construct and use that as a variable and look at its correlations with some sort of independent outcome or maybe an independent process measure.

And then the highest level of validity test time is when you actually have a direct measure of the quality construct or at least, most of it, and then you look at the association between that and an independent outcome or maybe a related process. To me, this is – of those three levels, this is at the lowest of those three.

David Cella: Well, it's a maintenance measure so the bar has to be higher than the lowest of the low, right?

Ashlie Wilbon: Yes, I'll just add – this is Ashlie again, that looking at their submission from last endorsement period, it looks like it was a new measure in the last review and so they just submitted face validity.

David Cella: It has face validity.

Matt Austin: Right, and this is Matt. I mean, I actually my experience with (STS) is they do a stellar job and develop great measures, but if I am having to assess it, they somewhat – they provided, I feel like it doesn't quite meet what has been asked of them.

Ashlie Wilbon: I think was it – forgive me, if it was Jeff that was saying that this is the lowest of the low, I guess of what you would expect to see. I guess the question for everyone would be, even though it's the lowest, is that good enough? And maybe consider that as you consider what you might make your final vote.

David Cella: I mean, I just – this is Dave – to speak for myself. I hear the troubling – the trouble, but I guess I'm willing personally for this to go to the parent committee or whatever we call the deciding committee for them to deliberate. It's a tough call though. This is, like you say, this is going to be carryover to the next one.

Ashlie Wilbon: Does anyone have anything else to add or do you guys feel ready to cast your vote at this point?

Jeffrey Geppert: This is process – sorry, go ahead.

Karen Johnson: I was just going to say …

Bijan Borah: This is Bijan …

Karen Johnson: … just to put it back into play for you. It does have face validity, but since it's a maintenance measure, you have to vote beyond that, so it's exactly what Ashlie had suggested is this graph that they did. Does that actually demonstrate validity to you empirically? I'm sorry, I heard Bijan …

Jeffrey Geppert: (Inaudible) proper clarification. If we voted low, then what's the result of that? It doesn't go forward or …

Karen Johnson: That's correct. So if a majority of you vote low or insufficient, perhaps, it might be the better way if you think that the testing that they've done is not enough for you, then it would go back to the developer and with your comment, and hopefully they would do some additional testing and bring it back.

Bijan Borah: This is Bijan, actually. That's what my question was. So if we were to ding it, I guess, what is the specific test that we would like to have then as the developers?

Ashlie Wilbon: So that would be maybe under the question at the end of each of the summary that we provided you under the feedback to developer – that question. I think there was a couple of suggestions in there about when he provided the kind of low, moderate, high examples of what might be expected, but if you have any more specifics around that, that would certainly be helpful for us and for the developer.

David Cella: What I'm thinking right now is sort of like this point earlier about certain method and results, so it seems like in that context, low ought to mean invalid,

right? Like, we have concerns that this measure is not valid and it shouldn't go forward.

I think that's not quite what we're saying; I think what we're saying is they just didn't demonstrate validity using the method.

David Cella: That would be insufficient.

Jeffrey Geppert: That would be insufficient.

David Cella: We can vote insufficient.

Jeffrey Geppert: Yes, but like what I am struggling with, is I'm thinking – they did face validity. They did some empirical validity testing …

David Cella: Yes.

Jeffrey Geppert: What I'm contemplating is kind of what you were saying before, is maybe what that means is moderate and it should go to the committee and a no bunch of caveats.

David Cella: Yes.

Jeffrey Geppert: Rather than trying to push – they are pushing to the curb.

Karen Johnson: Yes, so with the insufficient one, as you mentioned, it would basically – what you would say is that you don't have information you need to really determine if this measure is valid and that's a little of what you were saying earlier, but if you feel that you can kind – based on the method they used, well, if they are your favorite, it is something you can use and the results were in most consideration good enough to indicate the measure was valid, then moderately leading would be seen more appropriate.

And then, obviously, you can list your concerns and we'll incorporate it. It's already been mentioned and then the standing committee can decide, "Well, is this really good enough to keep going?" So those are kind of the different routes. You could go based on how you are feeling about information.

Ashlie Wilborn: Are you guys feeling like you're ready to vote? Now, on that measure, could we ask you to kind of pull out the SurveyMonkey for this measure in 2561 and cast a vote based on those guidance there on the insufficient versus the moderate?

David Cella: So it seems like if our inclination is with some amount of trepidation to move this forward to the review committee, it would be moderate – it would be the vote; and if our inclination is to say, "Here are some specific information that we don't have and need," we should be clear about what that is and vote insufficient and then we'll just see where the vote lands?

Ashlie Wilborn: Yes, I think that's an accurate characterization.

David Cella: All right, let's see what we think.

Ashlie Wilborn: You should be voting now, OK, and I'm going to move on to 2563, we've only got about five minutes; however, like I said this measure is very similar and the concerns raised were basically identical. This measure is 2563, it includes both aortic valve replacement and CABG, coronary artery bypass graft composite score. The composite score is constructed in the same way as the prior manager, specified at the same level of analysis, timeframe is the same. Same exclusions with the patient population and participants. They performed the same tests and again, on this measure, in terms of voting from the panel, it passed on reliability with one high and four moderate and for validity, there was a split of three high and two low.

And so, again the same issues are bulleted out here that we discussed previously, so I guess, with this measure, we would just ask if anyone had anything that they thought was distinctively different. I don't think that there are necessarily – potentially the actual score and values that were identified with the testing.

But let me just open it up for discussion while we have a little bit of time left and see if there is anything different that folks would want to raise with this measure and then see if you would be ready to vote.

David Cella: I don't have anything. This is Dave.

Ashlie Wilbon: OK, so it doesn't sound like there's anything distinctively different in terms of their approach to demonstrate reliability and validity, so Karen, unless you have anything else to add, I'm going ask that folks, just go ahead and vote on the validity for this measure keeping in mind the similarities between this one and the prior measure to make sure we're being consistent.

Karen Johnson: No, I agree.

Matt Austin: On just the validity or reliability and validity, both?

Ashlie Wilbon: Just validity.

Matt Austin: Why is that?

David Cella: Because reliability …

Ashlie Wilbon: Yes, it passed on reliability.

Matt Austin: Oh, I see. OK.

Ashlie Wilbon: No problem. And so while you guys are doing that, I'm actually going to pause and hand it back over to Karen with the caveat that we did not quite make it through to the last to measures. The good news is that they're both very similar and so, similar to what we did with these two measures, I think the discussion, there will be a lot of carryover between the two measures and hopefully, won't take too long, but we will need to reconvene on the next call that is scheduled for this subgroup to get through all the measures.

And I will go ahead and hand it over to Karen and the team to see if there is anything else to add in terms of next steps.

Karen Johnson: Thank you, Ashlie. This is Karen. The only thing I would add is between now and our next call, which I am sure the team will tell us, remind us when that will be. If you – we will give you one more chance on that next call to (pool) the other two, if you want to, it's absolutely fine if you do not want to (pool) those last two measures, but we will give you one more chance on that second call.

So Poonam or Miranda, May, do we have anything else or do we just want to bid goodbye and let everybody go?

Miranda Kuwahara: I think we can just give a goodbye. Again, our second call is on October 18th. I think it is around 2:00 p.m. as well. So we'll see you then and we can use the same link, right, the SurveyMonkey link for that next call.

Karen Johnson: OK.

Bijan Borah: OK, thank you.

Karen Johnson: All right, thank you, guys, so much.

Female: Bye.

Bijan Borah: Bye.

Karen Johnson: Bye.

David Cella: Thanks, bye-bye.

END