National Quality Forum

Moderator: Scientific Methods Panel
October 11, 2018
12:00 p.m. ET

OPERATOR: This is Conference # 9695827.

Miranda Kuwahara: Hello and welcome to the Methods Panel Subgroup Number Two Measure Evaluation Call. My name is Miranda Kuwahara. I'm with the National Quality Forum and I'm joined by my colleagues, Karen Johnson, Andrew Lyzenga and Ashlie Wilbon.

We'll begin with a roll call of methods panel subgroup number two members. First off, (Larry Glance)?

(Larry Glance): Here.

Miranda Kuwahara: Thank you. (Karen Troymatic)?

(Karen Troymatic): Present.

Miranda Kuwahara: (Mary Beth Varcor)?

(Mary Beth Varcor): Here.

Miranda Kuwahara: (Jean Nuccio)?

(Jean Nuccio): Here.

Miranda Kuwahara: And (Christie Splend)?

(Christie Splend): Hello. Yes.

Miranda Kuwahara: Thanks very much. So before we dive into measure discussions, I wanted to make a few housekeeping remarks.

A discussion guide was sent to subgroup members on Monday and this document will guide today's measure discussions and will follow the order presented on this document. Consensus was not reached for the first six measures presented on that discussion guide. And then, that – those measures will be slated for a discussion today.

All other measures received passing ratings and will not be discussed on today's call unless a member of the subgroup would like to take this opportunity to pull one of those measures to discuss them further. If you choose not to discuss – to discuss additional measures, the decision from your preliminary analyses will be made final for those passing measures. So I'll pause now to give you an opportunity to review that list of passing measures.

(Christie Splend): This is (Christie). I don't need to discuss them, but I just wanted to make sure that some of the issues that were raised about some of those measures, particularly the three or four returns to community – discharge community measures could be discussed at the standing committee when it convenes.

Miranda Kuwahara: Sure. So we will capture the sentiments you put forth in your preliminary analyses and we'll send that along in a package to the standing committee so they'll have your notes for those measures.

(Christie Splend): Thank you.

Miranda Kuwahara: All right. Well, hearing no other remarks from members of this subgroup, I think we should move on.

So in that same e-mail containing your discussion guide was a link to a SurveyMonkey. We ask that you pull that link up now and capture votes on reliability and/or validity at the conclusion of each measure discussion and the staff will prompt you on when to cast those votes.

Timing on today's call will be limited. We have about 18 minutes at each measure. We'd like to come to consensus on all six measures today. We do

have a follow-up – follow-up call scheduled on Wednesday, October 17, from noon to 2:00 p.m. Eastern Time to discuss any items we don't address today.

And then, finally, I'd like to note that this is a public call. However, there will be no opportunity for public comment and members of the subgroup cannot direct any questions to developers. For recordkeeping purposes, we do ask that you say your name before providing any remark.

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

Karen Johnson: Thank you, Miranda. And she covered just about everything that I wanted to mention, so thank you, Miranda. Just one note about the SurveyMonkey – we're asking you to vote when we direct you to mainly so that you don't have to remember to come back later and do it.

But we will – we don't have the ability to see your votes live, so we won't know what – what the voting results are until sometime after the call. So just wanted you to know that.

And I can't see the link. Can you point me to it in the discussion guide? I – is everybody else finding it because I don't see it.

Miranda Kuwahara: So it is actually not in the discussion guide. It's contained in that e-mail where the discussion guide …

Karen Johnson: In the e-mail, OK, got it.

Miranda Kuwahara: And if you'd like me to resend that e-mail, I'm more than happy to.

Karen Johnson: No. I have that. I just thought you said it was in the discussion guide. I've got it.

Male: Yes. It is – it is the link, right? It doesn't – OK. The first time I tried it, it gave me – it couldn't work, but.

Miranda Kuwahara: Let us know if you have any issues with that link and we'll try to troubleshoot from here.

Karen Johnson: The other thing, you know, we did try to do our best with this discussion guide. Hopefully, any feedback you have about whether you found it helpful or other things that you like to have seen, that sort of feedback we'd very much like to hear from you after the call so that we can improve as we go along.

You – I'm sure you noticed. What we try to do with pullout especially that either – I don't want to say the more serious concerns, but probably the concerns shared by more than one of you are things that we really thought really needed to be aired. That doesn't mean that there weren't additional things that you noticed as you were doing your evaluations.

So if you don't see them on the discussion guide as being something that we definitely want to hit on in today's discussion, feel free to bring those things up. So we just – we didn't – we didn't want to put every little thing on there, but that doesn't mean that we caught everything that really has to be discussed.

The thing about the voting, if you were in a meeting here at NQF and you're doing your voting, if a measure went down on reliability, then we would just stop discussion at that point and not continue the discussion for validity and not vote for validity, that sort of thing.

Since we don't – we won't be able to look at voting live and know where you went and where you landed on certain criteria for certain measures, we will discuss reliability and validity to the extent that we need to and that is listed in the discussion guide. So it may be they will end up discussing, for example, reliability and validity for a particular measure, but the measure may actually go down on validity.

No worries about that. It would still go down even if you had votes on validity and discussion. We will capture that discussion et cetera and share all of that with the various groups.

We will be sharing a summary that we're going to write with the standing committee. That's how – they're going to know about your concerns and your discussion and your ratings from – really from two ways; one is the summary

that we're going to write; and the second is we will also provide those preliminary analysis forms that you filled out. They will see those as well.

If a measure actually doesn't pass – and let me – let me rephrase what I just said – measures that pass will go to the standing committee and they will see both the summary and your preliminary analysis. If measures do not pass on either a reliability or validity, we will tell the standing committee, number one, that you discussed the measure; number two, that it went down on one thing or the other; and we'll provide a very brief rationale about why.

But we won't be sharing your detailed analyses with them. We will, however, share those with the developers. So hopefully, that makes sense, but that was a question that came up on subgroup one call, so I wanted to make sure that I mention that.

The other …

(Inaudible)

(Larry Glance): Karen?

Karen Johnson: Yes?

(Larry Glance): (Larry) – it's (Larry Glance). A quick question about process. So if I understand it, what you're saying is that if we do not pass a particular measure on one of the criteria, that discussion will not be shared with the standing committee? Did I misunderstand it?

Karen Johnson: You didn't completely misunderstand it. What we would tell the standing committee is that this measure came through; it was submitted; the methods panel looked at it; the methods panel did not pass it on, for example, reliability and a very brief discussion about why it didn't pass. For example, perhaps you didn't agree with the methodology that was used or perhaps you didn't think the results were strong enough or something like that.

So it would be – it would be a fairly brief treatment of why it went down. And then, the hope is, of course, that if something does go down that the

information you provided in the PAs as well as the discussion that you're having today will be great input for the developers to go back and add stuff perhaps to their submission, clarify things et cetera, and then bringing them back hopefully in the next cycle.

They don't have to bring them back in the next cycle, but we hope that's what would happen and then we do it again. And, if it passes that time, then the standing committee would give the full summary, the full PAs, all that stuff.

(Larry Glance): I see. So we're a hard stop then, essentially.

Karen Johnson: You – yes. You are a hard stop for reliability and …

(Larry Glance): Got it – got it. OK. Thanks.

Karen Johnson: Yes. The only other thing that we do acknowledge is that you may have to be switching back and forth between documents and we'll be doing the same thing. So you have your own PAs that you can look at.

If you want to pull up your colleagues' PAs and look at those, you have those as well. You will have measure submission forms that you might want to have handy as well as the discussion guide. So bear with us as we kind of pull things up and we'll do the same for you if we need to go hunting for particular things.

OK. With that, the way that we set this up is we split this out really between Ashlie and Andrew and myself in terms of measures that we would lead. So Andrew is going to lead us in the discussion of the first three measures and then, if we have time on this call, we get through all those, I'll take over and start with the next three that we are planning to discuss.

Ashlie (looked) out. Hers pretty much – there wasn't too much disagreement on her measures that she could certainly clarify things if needed. I will say – and it sounds like you've already done it – the discussion guide, even though we're not planning to discuss those seven measures at the bottom of the guide, we did try to put some things on there that may help you the next time you evaluate things, in terms of maybe our criteria or certain statements that we

wanted you to just know about. So we do hope that you read them even if – even if we don't (discuss them).

Towards the end of the call – and Miranda help me remembers – towards the end of the call, we will give you another opportunity to pull one of these seven if you need to. We don't want to artificially limit discussion if there's something that you really want to bring out, but you don't have to. So – but just don't let me forget to offer one more time at the end of the call.

OK. With that, I'm going to hand it over to Andrew to start us with Measure 2539.

Andrew Lyzenga: OK. Thanks, Karen. So 2539, this is a maintenance measure. I wanted to kind of highlight that upfront because I think there was some confusion about that.

It is a maintenance measure that's gone through our process before and is returning for re-endorsement. This is a measure of unplanned hospital visits within seven days of an outpatient colonoscopy procedure.

We did have a – we did not reach consensus on reliability among our reviewers. There was a consensus that validity could be rated moderate, but we had some sort of questions and concerns on – at the staff level that we wanted to discuss, so we'll talk about those as well.

To start out I guess with reliability here, we got – so the developers provided us a number of things and there was a little bit of confusion I think again with the reviewers on whether – which kinds of testing was provided.

They did suggest that they had provided some information around data element reliability. What they gave us was some information about previous audits that have been conducted of claims data and then they showed frequencies of the risk adjustment variables, I believe, over time.

Both of those kinds of analyses are not analyses we would consider at NQF to be testing of the measure at the data element score of reliability so that we would consider that insufficient in terms of data element reliability. But the

developer did also provide a couple of different kinds of score level reliability testing. And if the results of those analyses are adequate, that would suffice to pass the measure on reliability.

So we wanted to sort of clarify that there was, I believe, one or two reviewers who said that they wanted – they did not accept that this was a reliable measure because it didn't have sufficient data element reliability testing. But we wanted to make sure to highlight that score-level reliability is sufficient to pass a measure sort of by our policies and guidance.

So I think the sort of main question on reliability – again, sort of passing over to data element question and looking at the score level of reliability, they gave us two kinds of testing.

They did a split sample analysis and an interclass correlation of those results. And then, they did a signal-to-noise ratio using the Adams Method. They got fairly low reliability results for the split sample testing and then the results were quite a bit higher in terms of reliability for the signal to noise at the facility level.

We – let's see. Maybe I'll just open it up for any discussion we have at this point on reliability again looking mainly at the split sample and signal-to-noise analysis and whether those results were adequate by your judgment.

(Larry Glance): So this is (Larry Glance). I'm happy to start out the discussion. So I rated this as being high reliability based on the performance of the score.

I did that based on a signal-to-noise ratio which depending on which level the score was being analyzed was either 0.8 or 0.89, which, if you were to use a Lantus scale, which we can debate whether or not that's appropriate, but that would indicate at least substantial and nearly almost perfect reliability.

In terms of the other approach that they used where they did a split sample approach, that is a very conservative approach to looking at reliability and it's probably the lower bound. But I think since they provided both I think we can certainly go with the signal-to-noise ratio and it – I think it would be reasonable to pass it based on that.

Andrew Lyzenga: OK. Any other thoughts from the reviewers?

(Jean Nuccio): Hi. This is (Jean). I also rated the – score reliability as moderate to high and – but the data element was what I was concerned about, so I would concur with what (Larry) just said.

Andrew Lyzenga: OK. Anybody else?

(Inaudible)

Female: (Yes, this is) – go ahead, Karen.

Karen Johnson: I was just going to say I don't – if someone can help me understand why we get such (different) results between ICC and the other method because, if one is clearly better than the other, should we not just be requiring that everyone do that if the ICC is misleading in terms of reliability?

Andrew Lyzenga: I think (Larry) mentioned and the developer suggested this as well that the ICC or the split sample analysis is sort of a more conservative estimate of reliability. I'm actually not familiar enough with the methods here to know why that is or how that works and I don't know if (Larry) – if you want to …

(Larry Glance): I'm happy to comment on that. So, Karen, I will freely admit that I have not read the literature on why the ICC is a conservative estimate.

If you go back and look at the studies that have been done looking at reliability, the study in New England Journal by (Mahotra), and then the work that was published in Rand; they – I think they both used the signal-to-noise ratio as the primary method of evaluating score measure reliability.

And, to me, this – that particular approach is – has – it's very intuitive, so I tend to accept it. But I can't give you the statistical reasons why the other approach, the split sample approach, is more – why it's more conservative.

Karen Johnson: Fair enough. I was just curious as to whether there was a reason why they like (it so different) from each other. If one is always going to be much more conservative than the other, if I were a measure developer, I would not be too

(jazzed) about showing something that – it's (working) (inaudible) that's much less conservative, you know.

(Larry Glance): Right. Agreed.

(Jean Nuccio): Karen, (hello). This is (Jean). I was wondering if the interclass correlation of – is at a different level than the facility level that it – but I don't know how that might be at the score because the score is for the facility. So it's a little – it is a little bit confusing.

Male: The split sample is at – is at the same level, it's also at the facility?

Male: Yes.

Male: Essentially what you're doing...

Karen Johnson: Yes.

Male: Yes. So that they're being done at the same level.

Karen Johnson: They are. They are and I would just add that even though it's a conservative method, you know, ICC scores of 0.36 for the facilities and 0.3 are – you link three years of data, right – are really low, meaning your facility-level score even, you know, three years of average data could vary dramatically over time and that's really upsetting to facilities – organizations when they look really, really good one period and then look really bad the next.

So I – and it's really looking at the – how those scores are fluctuating over time which is, in practice, what people are looking at. So I know that the – the noise – signal-noise ratio is well accepted and it – and it – and it certainly indicates more power here. But I'm still concerned about these fluctuations in facility-level scores over time.

I also disagreed with their agreement as to why …

(Inaudible)

(Larry Glance): Can I (inaudible) comment on that?

Karen Johnson: Go ahead.

(Larry Glance): And I just – so the way – my understanding of the way we do a split sample test is you basically take your dataset with all the patients on the hospitals and you randomly split it into two halves.

Karen Johnson: Right.

(Larry Glance): And then, for each half – and you don't do this like in year one versus year two. You do it – so you just take all the data of all three years. You randomly split it and then you re-estimate the model in each of those and you compare the risk standardized, say, rates for whatever your outcome of interest is …

Karen Johnson: Right.

(Larry Glance): … at the level of the facility in one sample versus the other sample using an interclass correlation coefficient.

Karen Johnson: Right.

(Larry Glance): So you're not really looking to see whether or not hospital facility performance tracks over time. And again, talking with (Zenqui Lee) who is – who is the biostatistician at Yale and he's a guy on our – on the methods panel – he does say that this – the ICC is pretty conservative, so it's a lower bound. But the signal-to-noise ratio is really what most people, at least in the literature, what most people use.

Yale is really kind of unique about using that for all their measures. So I don't know that we should conclude based on a low ICC value that the measure is not reliable because it's really not the standard that most people use other than the Yale group.

Female: Yale then I go back to Karen's (point) – I mean – and I think it certainly could indicate that – I mean, if you take a random two samples and you're getting that huge of a difference in how you would evaluate how that facility performed, it feels like a problem to me. That's all. And you could see that

from year to year then certainly given these are big-enough samples and it looks like they would be with three years of data.

And the argument they gave though for this being a conservative measure is that, you know, this measure they said is more similar to assessing personality disorder where you could expect a lot of noise. Then, wait, what should be more consistent?

Yes, I disagree. Whether the person had a hospital visit seven days following a colonoscopy, it's pretty darn straightforward. It's – to me, it's not like, you know, I can – I can evaluate your personality a whole lot different than the next person. So I kind of disagreed with the – their validating argument as to the conservativeness, but just my two cents.

Andrew Lyzenga: Any other thoughts or comments? I think we'll probably want to revote on reliability as …

Female: Absolutely, yes.

Andrew Lyzenga: All right. Yes. OK.

Female: And as a matter of fact, if you have your surveys up and can go ahead and do it. It would be great if you could go ahead and just cast your vote for reliability.

Andrew Lyzenga: So we will see how the results of that come out and in the meantime we can move on to validity where we did get a moderate rating again from – for reviewers. But we wanted to mention that – again, reiterate that this is a maintenance measure and we did not consider any of the information that they provided as to be empirical validity testing.

They provided only eight validity testing for this measure. That is something that, for maintenance measures, we do not typically accept face validity unless the developer has provided adequate justification for that for not providing us empirical validity testing results.

Their justification was essentially that they could not find an appropriate measure with which to compare the results of this measure to sort of see if that sort of construct validity.

There are other ways that they could have done it, and so, we wanted to just sort of throw that back at you guys to see if that seems like an adequate or a sufficient justification for not doing empirical testing to you. If you do not think it's adequate justification, then we would recommend not passing this on validity because it does not meet our standard.

(Larry Glance): So I'm going to start off again if you don't mind.

Andrew Lyzenga: Sure.

(Larry Glance): When I think about measuring a validity, I think that there are three ways of doing it. The first way is face validity, the second way is construct or congruent validity. And I'm going to – and I'm going to throw a third one out which we talk about is predictive validity.

And the idea of predictive validity is you basically take the – so just going back, you can look at empiric validity using construct validity, so you can compare it to other credible measures. The issue with doing that is sometimes the other credible measures may actually not be any better or maybe worse than the measure that you're evaluating and we don't really have a gold standard to use to compare a measure to, so we can't really look at criterion validity.

In terms of predictive validity, what you're doing is you're essentially – you have a risk adjustment model for a risk adjustment measure. If you had a perfect measure, then you could predict with absolute certainty what would happen – what the outcome would be for an individual patient and then you could then predict the outcomes for all of the patients in a particular facility or provider and compare them to the observed outcome. So now you have a way, in a sense, a gold standard for looking at the performance of that particular facility.

So what predictive validity does is essentially it evaluates the performance or how good that risk adjustment model is. So I would – I would argue that predictive validity is a – is a way of looking at the empiric validity of a risk measure and so that when the measure developers provide us with evidence of, say, the discrimination and calibration of the model they are, in fact, empirically validating the model.

And so, based on that, I actually rated this as a – as a moderate for validity. And the reason I did that is because the C statistic measure of discrimination was certainly not great. It was about 0.66, but it was acceptable for this kind of a measure. Readmission measures typically have low values for C statistics.

And then, the calibration curve which I think is actually more important when you're looking at model performance for risk-adjusted outcome measures was actually very, very good. And that was why I rated this as moderate and, again, I – based on what I would consider to be empiric analysis.

Karen Johnson: So, (Larry), this is Karen from NQF – and I know that what you just said is something that we're – that we're working through in our draft of our paper.

But, traditionally, NQF – and it's something that we'll – we'll keep working on – traditionally, at NQF, we want to know that the risk model is adequate, right, so that gets to your predictive validity and your C statistics and your (goodness to say), that sort of thing.

We definitely want to know that. NQF has also said, in addition to that, we want other kinds of testing to show that the measure score is valid in this kind of cases for outcome measures. So we completely, today, agree with your statement about predictive validity and why it's important. But we think we need even more on top of that.

So the issue with this one is they didn't give us other besides what you just mentioned so that the question really is are you – are you OK with the justification that they provided for not doing additional testing and, if you are, then, as long as you're happy with the face validity assessment, then we're pretty much good to go on the testing part.

We still have to talk a little bit about the other threats to validity that all get wrapped into the rating that you give for validity. But hopefully that makes sense.

(Larry Glance): No, it does. Can I push back just a tiny little bit?

Karen Johnson: We can just a tiny little bit, but it might have to be something that we (thrush) out on one of our monthly calls if we're not …

(Larry Glance): OK. I'm just – it's just going to be a three-second thing pushing back. So I hear what you're saying and I totally get it and we need to do what the NQF does traditionally.

But I would say that, in my experiences on the standing committee on the – for the readmission measures – we – it seems like if people have made a pretty good argument that the risk adjustment model is pretty good, most people have kind of in a very subjective way voted to pass a measure on validity based on that without necessarily having to see other empirical analyses.

And again, that's it. That's all I'm going to say because I know we're tight. But I just wanted to comment – make that comment.

Karen Johnson: It will definitely be something that we'll talk about kind of offline at some point.

(Larry Glance): OK. All right. OK. I won't waste any more time. Thanks.

(Jean Nuccio): Yes. This is (Jean). Just a real quick question, is there some place on the form that the agency or the developer provides that could be more distinctive in terms of whether this is a new measure or a maintenance measure? I had particular problems on this one because …

Karen Johnson: Yes.

(Jean Nuccio): Because they only gave face validity I presumed that it was a new measure.

Karen Johnson: Yes. No, it's a really good question and something that we should have pointed out to you. On the measure information form, so that first piece that – the form that has all the specifications, at the very top of that form, there is a – what we call a brief measure description section on there. And towards the end of that section, it actually has a couple of fields that tell you the previous endorsement date.

So if there's a previous endorsement date, then you know (that it's not a new measure); it's a maintenance measure. If that is blank, then most of the time you can feel comfortable that it is a new measure. Every now and again there is something funky that goes on with our numbering system, so that's just kind of the FYI on our part. But that's how you would tell.

(Jean Nuccio): OK.

Karen Johnson: And we may need to consider is there a way that somehow on that brief measure information that we put that on there very bluntly so that people don't have to guess (and hand it back).

(Jean Nuccio): I was worried that if it was blank they might have forgotten to put it in.

Karen Johnson: No. I think that's actually a field that we put in on the form based on our records, so …

(Jean Nuccio): OK.

Karen Johnson: Yes. But it's – that's a really good point and you weren't the only one that got a little mixed up on a couple of measures as to whether they were new or might not. So that's a lesson for us that we need to make sure that that is more clear.

Andrew Lyzenga: So I guess getting back to sort of the core question we have here is whether their justification for providing only face validity was adequate, again, so setting to the side for now (Larry)'s point about risk adjustment being a form of validation. We – we're only going to consider right now the – their having provided face validity for testing. So is there adequate justification for that?

(Jean Nuccio): Given that this measure has been around for a while I found it strange that they would not have data from applying the measure overtime that we could use the – make some statement about validity.

Karen Johnson: This is one of those things that there is – there is really no right or wrong answer on justification. It really kind of depends on what you guys think sounds reasonable.

Sometimes justification for this kind of thing could very well be something along the lines of the developer saying, "Look, you know, we developed this measure a few years ago. We had money to do it. We have no more money to do additional testing. That's our justification."

And, that might be a reasonable justification. So there is – there's lots of things that might be and, as Andrew said, they didn't feel that they had a good – a measure that they could use for some kind of construct validation. So you can take that or leave that and make your own decision there.

I think – and, Andrew, I want to – I want to hand this back to you too, but the – your writing for validity won't only be based on the face validity justification. We also want to talk a little bit about meaningful differences and if there is any concerns with the risk adjustment approach. Let me hand that back to Andrew to talk about …

Andrew Lyzenga: Sure. Yes. So the question of meaningful differences, the developers provided some – the result of some analyses that showed by their sort of – I don't know if you would call it a ranking methodology or their method of identifying outliers – they found that only one of almost 4,000 outpatient facilities performed better than the national rate and only one performed worse than the national rate.

Eight hundred fifty-four of those close to 4,000 were classified as number of cases too small. In terms of the ambulatory surgical centers, none of them performed better than the national rate and only four performed worse than the national rate.

So that – given that when we endorse the measure through NQF we are implying that it is suitable for use and accountability purposes, that might seem like not a lot of differentiation, you know, with which to make sort of determinations of performance for accountability. Were there any concerns among the group about that?

(Christie Splend): No, I certainly have concerns about that. This is (Christie). You know, we really need to show that this rate can differentiate performance between providers so I really would have liked to see some empirical evidence showing that there is a distribution that there are good performers and bad performers and what they shared certainly doesn't show that to be the case.

So it's concerning I think. And I – and I get that these developers don't have a lot of money, but, if you approved it in the first instance based on face validity and now we're approving it again just based on face validity what people feel like it should show differences, but it's not, that's the concern.

(Larry Glance): So this is (Larry). I'll make a comment. So I think if you believe in the value precedent, when you're using these types of outcome where the incidence of bad outcomes is pretty low, it's pretty common not to have a lot of outliers.

And even for outcomes that are much more common, so for the CMS measures that are reported in Hospital Compare for pneumonias and AMIs and heart failure, they have very, very few outliers. So if you were going to nix measures based on the fact that they – there are – there are few outliers, you'd end up nixing a lot of measures that have already been endorsed by the NQF.

In part this is – the reason you have so few outliers is because you are using hierarchical modeling and shrinkage estimators and there's lots of reasons why people do that, not the least of which, it gives you some more stable estimates of hospital and facility performance.

But I think the fact that you do see a reasonable distribution in terms of the point estimates for the risk standardized rates I think is evidence that there is some variability. And again, given the fact that CMS is using a lot of measures where there are very few outliers and NQF has endorsed those

measures, I don't think we should nix this particular measure based on that finding.

Andrew Lyzenga: OK. Thanks, (Larry). Any other comment? Maybe we can briefly talk about risk adjustment as well.

There was some concern among our reviewers that the developers did not include dual status or any other social risk factors in their risk adjustment model even though their results – their analysis had shown, I believe, that the social risk factors may have some impact.

I should note that this is not a reason for the methods panel to fail a measure. This is kind of a little bit more of one of those clinical questions that the standing committees typically are well suited to weigh in on. But we can express our concerns as a panel to the standing committee and give them any thoughts that we might have on this issue.

Karen Johnson: So this is Karen. None of you will be surprised to hear that I have concerns with this decision. The dual eligibility was significantly related to the outcome with an (out) ratio of 1.5 for a hospital outpatient department, then 1.35 for ambulatory surgery centers.

And they actually specifically say that the effect (though small) does not vary across hospital outpatient department or ASCs in a specifically significant way. (They actually haven't) tested that, but that to me is sort of a definition of a factor that is (left) beyond the control if we see a very consistent effect across settings.

And they – and then, other measures justify the decision not to include it by saying that it doesn't change the model performance to substantially change measure scores which is not the – that's not a criterion that we hold other risk factors in the model to necessarily.

And, of course, things will be highly correlated with one variable in or out such a large sample and so many other variables. So I understand this is not a reason to fail a measure, but I am happy going on record to say that I'm still concerned with that decision.

(Christie Splend): And this is (Christie) and I agree with that 100 percent. Their data fully supports the fact that there are disparities. In fact, that very well might be the small outliers we're seeing (on an effort there) are really outliers, but the – certainly the performance of the rates was significant.

And, the argument after they have used the hierarchical model, so they are controlling for quality between hospitals or centers, they – so they can use the hierarchical model, they have a lot of other risk factors in there as you said, Karen. Yes, you wouldn't expect the – one dual status variable to affect the C statistic of the model itself. But it's certainly there is strong evidence of the fact that duals have – have more of these outcomes even in – even within the same hospital where you would expect the care to be similar.

So, the argument that it could be associate disparate care due to associate demographic status is – I think they did account for that with their hierarchical modeling and all of the other risk factors they included, so I don't buy it.

Andrew Lyzenga: OK.

Karen Johnson: OK.

Andrew Lyzenga: Well, I think then the next step is for us to revote also on validity. Is that right?

Karen Johnson: Yes. We are asking now I think – NQF actually (fooled) this one, so we would like you to vote, you know, since we pointed out that they really need that justification for testing.

Andrew Lyzenga: So again, and probably the things you'll want to be thinking about on this vote are that question of whether they provided adequate justification for providing only face validity and then this question of meaningful differences and can the measure adequately distinguish performance across providers.

So we will ask you to revote on both reliability and validity through your SurveyMonkey instrument. It's probably best if you can do that as soon as

possible while it's fresh in your minds. We can get those results as quickly as possible as well. So, with that, I guess we will move on to the next measure.

(Larry Glance): Quick process thing.

Andrew Lyzenga: Sure.

(Larry Glance): It seems like after you vote you have to go back and – for the next measure you have to go back and hit the link the again. Is – it takes you out of SurveyMonkey after you voted on one measure.

Female: That's correct. You'll have to go back and re-click on that original link.

Female: Or just copy the address and paste it into open windows.

Andrew Lyzenga: OK. Everybody all right with that? OK. So the next measure – and the next two measures I think are actually fairly similar at least in the concerns that were identified, so hopefully maybe we can get through one and then kind of quickly move through the next based on that conversation.

This is a measure of hospital visits that we're talking about, 3366 now, hospital visits after urology ambulatory surgical center procedures. So in terms of our concerns here, we did get a reliable – moderate rating on reliability. So I'm trying to refresh my memory here.

So again, we have here a reference to audits of claims data, I believe, that developers (say) generally demonstrate that claims data are reliable. The data might not provide us any results of those – of those audits and also showed us the frequency of risk adjustment variables over time.

Again, those two kinds of analyses are not things that we would consider data element reliability testing, but they did, similar to the last measure, do the same kind of thing with the split sample analysis and then a signal-to-noise analysis hitting fairly similar to the last one, a little bit higher split sample results, 0.45 here, and then around 0.7 for the signal to noise.

So let's see. The – this one. Yes. So maybe we can discuss reliability very quickly here, particularly that score level reliability that they provided. Any thoughts or discussion about that?

Karen Johnson: This is Karen from NQF. I feel like it's kind of the same conversation that we've just had for the other measure. For this one, since we had one high, three moderate and one low, we really are thinking that we don't have to revote on that one unless you guys feel that we should. Am I being correct on this?

Andrew Lyzenga: And I think the main – again, sort of – this is for learning purposes, clarification that we wanted to add here is that, given that they have provided measure score reliability results, they do not have to provide data element reliability. So just kind of wanted to clarify that for the group and that's all.

And, with that, unless there's any concerns about the reliability, we can just accept the voting results that we already have and move on to validity. Any objection to that?

Female: No.

Female: No.

Andrew Lyzenga: OK.

Male: No.

Andrew Lyzenga: All right. So, validity, we again have a similar situation here in that they have provided only face validity. But this is, in fact, a new measure so that is acceptable by our sort of standards.

The – let's see. Let me refresh my memory again here. I think we did get a consensus not reached on validity here, so they provided face validity here. They – and we have some questions again about meaningful differences. That same outlier analysis suggested that only 19 of the roughly 1,200 ambulatory surgical centers were better or worse than expected although again odds ratios

and distributional statistics did show some more variation in results across facilities.

With risk adjustment, again, same concern here; did not include dual status or any other social risk factor despite showing some impact on the results. And then, some concern with the C statistic that's reported for the risk adjustment model whether that was, in fact, adequate performance for the risk adjustment model.

So that's kind of the – overall the concerns here of points of discussion. Any thoughts on any of those issues either the testing, the meaningful differences or the risk adjustment?

Female: No.

(Inaudible)

(Larry Glance): Can you …

(Jean Nuccio): A quick question for my colleagues. This is (Jean). What does work relative value of units mean? I'm not familiar with that term. That was one of the variables in the model.

Andrew Lyzenga: Yes.

Female: The RVU (inaudible) a measure of – well, of value. So the – they are determined by a committee that sets how many work units each thing is worth, so a primary care visit versus a specialist visit versus a bypass surgery at RVUs (and so you're) – you pay based on RVUs.

I think I hear it's used to separate the complexity of procedures basically into something that it would be high versus low. It would be like a more complex versus less complex procedure.

(Jean Nuccio): OK. Thank you. I just – I couldn't find it in grey searches and …

(Larry Glance): If you – if you do – if you look at the older ACS NSQIP risk models they all used work RVUs as a measure of surgical complexity. So if you're looking for a source on that, you can go there.

(Jean Nuccio): OK.

Andrew Lyzenga: Any other discussion about validity.

(Inaudible)

(Jean Nuccio): So I have a quick question.

Female: Go ahead.

(Jean Nuccio): This is (Jean). If we're (deeming) this on the basis that it doesn't meet the NQF requirements or face validity, can you in 60 seconds tell us what those criteria are?

Andrew Lyzenga: So this one actually does meet our requirements given that it's a new measure, not a maintenance measure. Because it's a new measure, we will accept face validity …

(Jean Nuccio): It says here in your report that it does not meet NQF's requirements for face validity, unless I'm reading that wrong. Under "Items to be discussed, face validity. The other avenues of face validity described by the developer are fine, but do not meet NQF's requirement for face validity."

Andrew Lyzenga: I think – so I think that was because they provided us like a few different pieces of information, some of them we did not consider meeting our requirements face – for face validity. But one method that they provided, that being the (TAPT) review did meet our requirements for face validity.

(Jean Nuccio): I see. OK. Thank you.

(Christie Splend): Yes. And I was just going to comment that the same comments I think that Karen and I expressed before about not including dual status applies here, I think even stronger because I'm not sure whether a sample came from, but

their – they considered when they were comparing the scores, and so, there's not a big difference in scores.

They use 1.9 percent duals as low percent duals and 7.5 percent or higher duals as high duals. I can imagine that some of these centers serve 75 percent duals or 100 percent duals and their scores would be dramatically different especially given that even in – given the – their distribution, when you're just looking at (courtship), I mean, dual places tend to be – it – facilities tend to be the – almost no duals or almost a lot of duals.

So when you're just looking at quartiles, yes, to make the cutoff for the fourth quartile at 7 percent duals is really low compared to what the real world probably looks like.

So I'm not surprised they didn't find variations given the way they cut the sample and – but they did find, again, a significant gap in rates even given the low percentage duals that was considered a dual facility, you know, 7.5 percent compared to 5.9 percent for non-duals. So, again, argue that the decision not to adjust for that is – I disagree with.

Karen Johnson: On page 24 of the measure testing document there is a graph, the (scatter) (inaudible) the proportion dual. I don't know why they only show the top quartile and there's no relationship shown within this quartile. But, to your point, there is at least one facility that (sits) what appears to be 98 percent dual and actually a whole bunch between whatever that's the cutoff of these seven or whatever and 40.

(Christie Splend): Right. So those lower percent dual facilities, you know, that are in there and there's a lot of them are probably not going to be very much different from the almost no dual facility.

Andrew Lyzenga: OK. Any other discussion about validity? If not, then I guess, again, we will not vote on reliability for this measure. We'll just sort of move our previous result forward, but we will take a revote on validity, so we would ask you to do that again in your SurveyMonkey instrument.

Female: (Inaudible).

Andrew Lyzenga: Yes. Sure.

Female: The only concern about validity is about the risk adjustment with duals. We've concluded from voting, but should we not vote low and just vote moderate but, like, file our concern?

Karen Johnson: Yes. So what we've kind of instituted is we would like the methods panel not to take down a measure simply because you don't agree with the inclusion or non-inclusion of particular risk factors. We'd like that to be a discussion point for the standing committee. So we definitely want your concerns, but we don't want you to vote low solely because of this.

Female: You're killing me, Karen.

Karen Johnson: Yes. It's a little tricky because it's partly because of how we're setting this up to where the methods panel could basically kill a measure and, if that happens, it doesn't go to the standing committee. So we don't want to take that specific discussion away from a standing committee.

So if that really is the only concern that you would have, we want the standing committee to be able to weigh in on that as well. But we will definitely make sure that they hear what you're saying.

Female: Thanks.

Andrew Lyzenga: OK. I wonder if maybe we could – can we take that (discussion) and apply it to the next measure or should we talk through it or maybe we should just open it up to see if there is any additional thoughts on the next measure. This is I think a very similar one, 3470, so hospital visits after orthopedic ASC procedures.

I believe we have the same issues here with – I guess we had different voting results. So we had a consensus not reached on reliability and then a moderate on validity although I think we have very similar issues to the last one. We, again, have no data element reliability testing at least as far as NQF's requirements are concerned, but we do have testing at the core level.

Again, you know, split sample signal to noise. Any discussion of those results that we have for score-level reliability for the measure 3370?

Female: So this is (inaudible) in this case the ICC was 0.25 and the facility-level reliability or the –(whatever the other one), that was 0.66, which is not nearly as good as the prior one of 0.8, whatever it was. So, (Larry), how do these feel to you? Do you just trust the higher one or do you think the truth is somewhere in between? Or how do you think about when we have both of these things reported for the same measure?

(Larry Glance): So great question. I would go to the – with the signal-to-noise ratio because I think that's what has been reported in the literature as what people use. So I would go with the 0.66 which to me would indicate substantial reliability.

Andrew Lyzenga: OK.

(Jean Nuccio): This is (Jean). Is this another one where, because we have information about the score reliability, we don't need information about the data element reliability?

Andrew Lyzenga: Correct. Yes. So if their – if the score-level reliability, if you consider that to be adequate, then say, give it a thumbs up on reliability. Any other discussion about that? If not, I think because we had a consensus not reached on that criterion, we will revote on reliability for this one.

Validity, we wanted to talk about because, again, new measure, so face validity only is OK in terms of the method. We did have some concerns about how they conducted that face validity for this measure though.

We would like them to explicitly address the question of whether performance scores resulting for the measure as specified can be used to distinguish good from poor quality. They did not explicitly address that specific question.

They asked whether the measure that's specified is a valid and useful measure of orthopedic surgical quality of care and whether the measure specified will provide ASCs with information that can be used to improve their quality of

care. Kind of similar concepts but not quite the same thing as we would like them to ask.

And then, one other concern is that we expected that if there is disagreement among the (TEP) members that they would provide a reason for that disagreement or some discussion of that disagreement and they did not do that in this case. So …

(Larry Glance): This is (Larry). Can I make a comment?

Andrew Lyzenga: Yes.

(Larry Glance): So with – (it provides) of the comments that Karen Johnson made earlier, I would like, if possible, if I have a little wiggle room here, for us to take a slightly holistic approach to looking at measure validity. And, in this case, I get it that the – that the (TEP) did not quite meet the requirements.

I would argue that it would be very difficult to evaluate construct validity because there really aren't a lot of measures out there, if any, that look at performance for any kind of ambulatory surgical centers. You can't really look at hard outcomes because they are just so, so incredibly uncommon, so people look at these types of measures like – a defined for hospital visits.

So I think it's going to be very hard to look at construct or congruent validity. And I would still argue that predictive validity here is acceptable. So if you take the whole – the whole thing put together – I would argue that this is a valid measure. I would also argue that we should not, just based on the face validity, nix this measure.

Because, as Karen explained earlier, this is kind of a hard stop and that means the standing committees don't have a – the opportunity to evaluate this. So I would vote to pass this on to the standing committee even though it may not really meet the NQF criteria the way they are currently designed.

Karen Johnson: And this is Karen from NQF. I think we wanted to be careful on how we phrase this. They did do a systematic assessment of face validity. It doesn't

quite mean exactly what we want, but it should be that you guys think it's even close enough to what we want.

So even putting aside the predictive validity that they did with the risk model, you guys might look at the face validity assessment that they did and say, "Hey, it hits the spirit of what we're wanting and I'm happy enough with that." And if that's how you – if that's your decision, that's absolutely fine.

And I completely agree with (Larry) that when you rate validity, you have to take all of these things into account, so it's not just the testing. It is the adequacy of the risk model. It is meaningful differences. It is how – are the exclusions appropriate et cetera. So there's a lot in there that you have to kind of weigh and take into account when you vote on validity.

Andrew Lyzenga: Any further discussion? Again, we'll – I'll just say that similar concerns again here about the ability of the measure to distinguish meaningful differences between providers and about inclusion of social risk factors in the risk adjustment approach. But I assume we've got the same concerns again here and same thoughts about the risk – social risk factors should have been included.

Female: Yes.

Andrew Lyzenga: Yes.

Female: Yes.

Andrew Lyzenga: OK.

(Larry Glance): That's unanimous.

Andrew Lyzenga: Yes. All right. OK. So then, we will, again – we'll – so we will do a revote on both reliability and validity for this one using your SurveyMonkey tool. And, with that, I think we can move on to the next measure. Are we clear with those or …

Karen Johnson: Yes. I think we're mostly clear. I do have one question and I only want to spend, like, two minutes on this. But the whole thing about meaningful

differences and, in each case, they did an analysis and compared to the national mean.

Is that a reasonable approach or could you think of other approaches besides just providing the distributional statistics that might be useful? I'll just stop there and see if anybody has anything that you might want to suggest that people think about in terms of analysis the next time around.

OK. I guess not. If something occurs to you, feel free to let us know. We can always come back to that.

(Larry Glance): So, Karen, just a quick comment. I think that this is appropriate. I think showing the distribution of risk standardized (rates) is a great way to get a sense for how much variability there is, I think also categorizing them in terms of outlier status.

There's one more summary measure that people use, but it's kind of nerdy, but it's called the median odds ratio. And it basically compares your – a particular facility's performance to all the other performance. But you don't really see it very much in all the literature. So it's nice because it's a summary measure, but it's not so nice because nobody ever talks about it.

Female: It's used a lot in the cardiovascular literature for – in particular coming from a lot of registry work.

(Larry Glance): OK.

Female: It's – I've certainly seen it a lot more in the last year or two than ever before that, so I agree with you. This might be something worth looking into.

Karen Johnson: And actually, Andrew just found that for measure 3366, which is the first one, and maybe the other ones as well, they actually did report the median odds ratio. I had never actually seen that reported so I didn't know what to make of that.

So it sounds like it's a good thing to show. I don't know how to interpret a median odds ratio of 1.27. I don't know what that means. But maybe we can delve into that in one of our monthly calls.

Male: (Sure).

Karen Johnson: Yes. They did seem – they did report that on the …

Andrew Lyzenga: 3370, I think.

Karen Johnson: On 33 …

Andrew Lyzenga: Or 3470.

Karen Johnson: At least one or the other, so probably I think these were the same developers. So if they did it for …

Andrew Lyzenga: Right.

Karen Johnson: … one, they probably did it for all three. OK. Thanks for that. OK. We're doing actually quite well on time. We're a little over halfway through our call and we are halfway through our measures. So let's just kind of go to the next one.

The – actually, the next two are going to be very similar. We – I would like to discuss them separately, but the approach et cetera is pretty much the same in both of them. So looking at measure 33 – sorry 3443, all-cause emergency department utilization rate for Medicaid beneficiaries with complex care needs and high costs.

So this an ED utilization rate. The target population is the adult Medicaid – adult Medicaid beneficiaries who are these high-needs folks. So it's a fairly – in a way, one could say that it's a fairly narrow target population.

So in the discussion guide we talk a little bit about the definition that they used for these BCNs as they call them. And we added a little bit of information about the (current) condition warehouse. All of you probably are

familiar with that dataset, but if you're not, some of the information is there. And that just kind of gives a little bit of a flavor of the data that they used.

The – one of the things that came up I think that was confusing for this measure is that no denominator exclusions were listed in the submission. So that kind of tore people up.

So it is a new measure, I forgot to mention that. And it is paired with the admission measure for the same group, so two measures, one for ED and one for admission.

The – for reliability, there really wasn't too much disagreement between you guys. We ended up with three high, one moderate and one insufficient rate. And unless you guys feel otherwise, we would put this forward as passing with a high rating, just kind of going with the majority there.

We do point out that these – this measure was tested using (max) data. That is also a dataset that is probably a little bit less commonly used. Some people have used it, but maybe not everybody.

They actually develop the measure using data from 10 states, which sounds like not a lot of states, but they actually – I don't believe has (max) data for all 50 states. I think they maybe only have it for around 15 states and the developers actually, as a bit of an appendix or some extra information, did provide some information about what was available for them and why certain states they ended up not using it. So hopefully you had a chance to look at that.

There was a question they did signal-to-noise ratio analysis reliability. Their average reliability was 0.92, the range was between 0.59 to 0.99 across those 10 states. There was a little typo in there, but should – the overall SNR shouldn't be 0.99; it should be 0.92. So that's just a typo.

There was a question from one of the panel members about why a couple of those states had the really low reliability. Our assumption is that these two states are the ones who had a very low sample size. I don't know if that's

something that you guys agree with or not, but that's kind of what we were thinking.

And the only thing that was a little tricky about this measure is, in terms of exclusions, dual eligible beneficiaries are not formally excluded from this measure, but they were not included in the testing. And my understanding is because they didn't have the data available in the (max) data at the time to do that. So the idea is that they would be included if the measure is implemented, but they did not include them when they tested the measure.

So let me stop there. Again, unless there is a desire to, we won't do continued discussion about reliability. We will go with the high rating. So does anybody have an objection to that?

(Christie Splend): No objection. This is (Christie). I just want to make a quick comment that I had a – I was a little concerned – and I guess this is not – I rated at moderate because the scores were so high.

But, they didn't discuss the reliability of the data elements or the – or the risk adjustment models. I think like it should be really important for this measure because of the probable differences in the quality of Medicaid data across states and we know this to be true and that – and that wasn't tested.

So that's a little concerning to me and I just want to point that out. I think that should be brought to the steering committee.

Karen Johnson: Thanks, (Christie). As with the other measures that we've talked about today since it is a – kind of simple, quote/unquote "outcome measure", NQF doesn't require that they do both kinds of testing. So it's not that they didn't meet our requirements in terms of what they have to do, but I think it's certainly fair …

(Christie Splend): I think, yes.

Karen Johnson: Yes. It's certainly fair for you to say, hey, given the data that this is coming from and the – and the vagaries of the different data sources in the future if it goes through you would highly desire to see that type of data. Is that a fair summary?

(Christie Splend): Yes. Exactly. I feel like, yes, given the quality of Medicaid data differences across states that that's something that should be a concern when we're comparing the outcomes across states.

Karen Johnson: OK.

(Jean Nuccio): Hi. This is (Jean). I think my comments related to that, when NQF endorses a measure, are they not endorsing the measure for national use? And so, the question then becomes how could that be if there are only 15 states that have data or maybe more states have data, of which only 10 were used to develop it? And then, how representative of those 10 states, if, in fact, you could find data across all 50 states, of the other 40 states?

Karen Johnson: It's a really good question. So NQF does kind of endorse to some extent, like you said, nationally. Exceptions might be when measures or the data source is something along the lines of a registry.

So the understanding is that it would be applied to facilities or clinicians who submits a registry and possible, you know, would be kind of expandable to other types of people, but it's really endorsed for a registry dataset. This measure, that's not quite the analogous thing. The idea is that it should work for all Medicaid high-risk (benes), but they weren't able to actually build the measure using all of that data.

So that is a question really I think of validity that we want you to think about. Even though it feels like it's a sex question and it kind of is, it's probably more of a validity question. And I think you'd have to think about it in terms of do you feel like 10 states worth of data you can get a good risk model, you can build a good measure when you're limited to not all the states.

(Larry Glance): I – this is (Larry). I got a couple of comments. The first one is I think that's clearly a limitation when you're just using a sample of states and not all states, but that's a limitation that a lot of different models have in common.

So, for example, for years, people considered the New York State Cardiac Surgery Model to be one of the gold standard models for cardiac surgery outcomes and it was based on just one state's worth of data.

The second comment is with respect to the fact that the Medicaid data may be heterogeneous across different states. I think that's a – that's a really valid criticism. But I would say that you could make exactly the same criticism when you're doing hospital-level risk adjustment as opposed to state-level risk adjustment, meaning that some hospitals are much more aggressive about coding outcomes and complications than others – problems with coding.

Despite that, we still – we still use those datasets for performance measurement. So although it's a valid critique, I don't – I don't think it's a – it'd be something that we (all) shoot this measure down.

The one – the thing that I was concerned about and maybe I didn't quite understand this, but – and I think somebody else may have mentioned this in their review – but you may have different levels of churning of Medicaid recipients across different states. So at some states people may move in and out of Medicaid more quickly than others.

And I didn't have – and so, the question is, if – I mean, you're going to be excluded if you don't have 11 – I think it was 11 months' worth of Medicaid data or something – I don't remember whether 11 months or 12 months.

So, in some places, if you're in and out of Medicaid maybe you have a lot more people who are being excluded in certain states than other states and that could be a source of bias for the measure.

And I was wondering, Karen, if you could help me with that. I mean, would – I didn't get a sense for how many – what percentage of patients were being excluded from different states because of churning.

Karen Johnson: Yes. And I don't know – this one was a little bit tricky because it's – they really needed I think a 24-month lookback. So 12 months or at least 10 months for the measurement here and then a year prior to find a denominator.

So I don't know. I can't remember if they did an exclusion analysis that would give us that. This concern actually came from one of your colleagues. So I don't know what – if – whoever discussed this churning.

(Larry Glance): So I didn't use the term churning, but this was one of the concerns that I've put down in my review.

Karen Johnson: OK.

(Karen Troymatic): It was me. This is the other (Karen). That was me that put that down. I think with Medicaid – and this goes back to the prior point about Medicaid data – Medicaid is a fundamentally state-held program and data are very, very different.

And, for example, the use of managed care versus fee-for-service Medicaid; the quality of the data; the degree of auditing; the way things are paid. Missouri Medicaid can't pay by DRGs because they don't have a sophisticated and updated system to pay by DRGs. They pay cluster (cost minus). I mean, there's a lot of like really surprising thing about Medicaid data.

Also, eligibility varies from state to state, right. So your denominator population actually is quite different in different states and it's not just the stated generosity of Medicaid, but it's also all the administrative things that make it easy or hard for people to access or be the access from the program.

And so, it's not to say that that should necessarily kill the measure, but it is to say that I think that the developers really do need to provide some analysis of all the folks that are getting dropped and some comparability between states to help us understand what that looks like.

If we're just seeing a piece of the puzzle in each state, I'm not sure that it would generalize even to those states let alone much more broadly as we've already said since many states were not included in the testing.

Karen Johnson: So, (Larry), did that help and …

(Larry Glance): Yes. That really – that really helps a lot. So I guess the question is then in terms of validity because that's really what it boils down to. (Karen), what's your sense? I mean, how would you rate the validity of this measure based on the things that you've just brought up?

(Karen Troymatic): So I – when I did it, I sort of waffled and said moderate assuming that we could see that those exclusions were similar across states. I don't know that I – I guess maybe I should have said insufficient, so that it could be a non-issue.

I mean, we could be seeing a very representative sample and so it's included in the measure and it's such a basic, straightforward measure in terms of a big population, a relatively easily measured event. It may be that all this stuff makes no difference. I just don't know without seeing it.

(Larry Glance): I wonder if – because this is not being – the intention of this was not for pay for performance or reimbursement stuff, it's more just sort of to get a sense for how different states are performing.

Again, this is being – I'm going to put my very holistic hat on it. Maybe we should give them a little bit of benefit of the doubt on this because it would be interesting to get this type of information out even if it's imperfect as all measures are imperfect. I don't know what else people …

(Jean Nuccio): Yes. Related to that, what do differences between states mean, which is my concern on the validity area?

(Christie Splend): Well, one of the things that I had noted in my review was that they kind of use the NQF to say the NQF says that including area level socioeconomic indicators is not appropriate; it doesn't improve the predictive, you know, capacity of models for hospital-based care measures. I don't think that was NQF's conclusion and I think you said that in the – in the summary report.

But, in this case, certainly, state-level variables like disparities in income across state and poverty and education levels vary significantly. We don't know what – what states these are, but we can imagine that that's the case.

So we don't know if these results are generalizable or not purely because they haven't control for some of those just geographic differences and what these – and I – (Karen), you basically said that. These populations are going to look very different, but that could have control for some of that but they said that the NQF said that that wouldn't matter, so.

One more comment. The other thing they left out was (colleague) pharmacy just because it wasn't in the – in the – in the model or in – available in the data. And, given – like, for example, the current opioid epidemic and other issues around medication use and high-risk medications used differently across states especially in the Medicaid population or disabled population, you know, that also could really skew the results here.

Karen Johnson: And this is Karen from NQF. So a couple of other things. You've hit I think a couple of threats to validity that we kind of listed under the – under the (test).

(Inaudible)

Karen Johnson: Yes. But we also just wanted to point out that they did do face validity, it is allowed. What – the process (met) requirements that you would just have to look at that and see if you think the results are adequate. The 10 states – I think we've heard from you about the generalizability, if you have any advice about need for recalibration or that sort of thing.

There was also a couple of questions just in general and I'm assuming that these weren't fatal flaws from whomever voiced these concerns. That there wasn't a discussion about overfitting or the risk of overfitting, a little bit of question about factors with the protective effect included, and also just not understanding why child was in the model if this measure is limited to (benes) who are 18 to 64, so that's it's a clarity question there.

And again, thank you, (Christie), for pointing that out. We want to be very clear on this call that the most recent admissions/readmissions project report where several measures were endorsed without social risk factor control, we do not think that that means that you shouldn't be controlling for those in other measures.

We absolutely think that you need to decide and look into whether this conceptual rationale or a potential inclusion and then try to look at the data to the extent you can and do the empirical analysis. What you know about one measure and one population doesn't actually say anything about inclusion yes or no for another measure. So I want to make that point very strongly.

So with all of that, I think those were the concerns on the measure. So it sounds like (Larry) may be willing to say, "Hey, we know this isn't perfect, but it's probably a good start. It could really be useful."

There is – but also along with that that concern that, hey, it really was only using 10 states and we don't really know completely how those states varied and then also the concern about risk factors, particularly for this population.

(Larry Glance): Yes. I was the one who critiqued the risk adjustment model and it was – it was kind – I thought it was a little sloppy. They included a whole bunch of co-variants that were neither clinically significant nor statistically significant. So it was sort of like they had small (effect) sizes and the P values were huge.

So that didn't really make a lot sense. And, as you said, they included risk factors that seem to make it that you would have an advantage of – in terms of the outcome if you had this particular disease stage. It really doesn't make a lot of sense. Having said that, I still think that we ought to kind of give them a pass and let the standing committee take a look at this because I think it is an interesting measure and it may be something that people want to look at.

Karen Johnson: OK. Any other thoughts on validity? Did we (trash) out the major concerns?

(Larry Glance): Yes.

Karen Johnson: OK. So, with that, we would like you to go ahead and vote on validity for this measure. We're not going to ask you to vote on reliability. It doesn't seem like you guys had a problem with going with the rating – the consensus rating on reliability. Now, the good news is – I don't know how long it takes you guys to do that, but hopefully, it does a real quick little thing for you to do.

The next measure, 3445, is almost identical in terms of what it's trying to do. It is paired. It is the admission measure that's paired with the ED measure.

One thing that I thought was interesting to point out is that it is – it includes admissions and observation stays in the numerator. Once again, there was confusion about whether there are a bunch of exclusions in this measure or not.

And, that's something I really didn't really talk about too much before, but different developers sometimes call things exclusions that maybe I would not and I would just say that was a part of your target population. So sometimes the exclusions are a little tricky.

But we do know that the Medicare (benes) who have left the 10 months of data in the – in the lookback period are excluded from the measure. So the idea is there's not enough data there to be able to put them into the denominator population.

So for reliability, basically people landed in the same range. So we are going to put that one forward as passing with a high rating unless there is objections from you guys. However, on validity, again, there was kind of a spread in terms of rating and we wanted to discuss on the call.

So exclusions we've talked about in terms of validity testing, this one is a little bit different in the testing because they did do a construct validation, but they correlated it – the results of this measure to a (Keytas) inpatient hospital utilization measure that targets Medicare and commercial enrollees.

So the concern here was that, yes, they did do construct validation, but the populations were different. And what does that mean? Is that actually a good way or a reasonable way to validate this measure? I think we had the same things in terms of risk adjustment again relying on data from 10 states. (Larry)'s questions about the control overfitting et cetera.

And then, the threats to validity I think are pretty much the same that we just discussed in terms of the churn, in terms of would (be liking) more data so that you can understand the differences in the – in the states. And then, there

was an additional note about, of the 10 states, only three were different from the mean, so getting to that 10 meaningful difference question.

So I think we've probably discussed in the last measure almost all of this. So the thing that remains to be kind of discussed a little bit is if the construct validation with the Medicare/commercial enrollee hospital measure and is that deemed OK. Is that a reasonable test for you?

(Larry Glance): Karen, for – again, for the Medicare measure, can you – can – what is the title of that measure one more time?

Karen Johnson: Let's see. I think they called it (HITUS) inpatient hospital utilization measure. I don't know if that's an actual – that's probably not the formal title, but I think that's all I have.

(Larry Glance): And it's also at a state – it's also a – they're using it as a – at the state level when they did the comparison?

Karen Johnson: Let me see if I can bring up the testing attachment. Bear with me just a second.

(Larry Glance): Sure.

Karen Johnson: And if anybody beats me to it, that's fine too. Sorry, I thought I had them all (listed here).

Female: Karen, I think the (HITUS) measure is similar. It doesn't actually say kind of what – where the (HITUS) measure is calculated. It does say state-level performance of a conceptually-related quality measure. It doesn't say whether or not they aggregate it or whether or not it is available already aggregated.

So it's the same concept, but a slightly different population, Medicare and commercial enrollees; and a slightly different specification in medical and surgical inpatient events and excludes inpatient events due to behavioral health clinicians and maternity.

It basically suggests that hospitalization rates are more similar or highly similar by state across payers which is I think consistent with my understanding of prior work on geographic variability, right.

Karen Johnson: So I don't recall who had the concern about this particular measure being used in a construct validation. Is that some – if you guys remember who it was that pointed it out, did you feel like that was a fatal flaw or did you feel like it's just something you wanted to point out that …

(Mary Beth Varcor): This is (Mary Beth Varcor). It was me who pointed that out and I did have some concerns only because the populations are different. But as far as a fatal flaw, I'm not so sure that that's, you know – I waffled, so I'm still on the fence about it. I'm not so sure.

Karen Johnson: OK. Thanks. I'm not sure if there is much more we can talk about here. I think you probably just have to think about it a little bit yourself and see if you feel like that is maybe a reasonable effort for a new measure. And I'm not aware – I mean, there are – I mean, I'm kind of surprised they didn't correlate it with the ED measure. That would have been kind of the trivial solution that I think they could have done.

OK. Does anybody have any other comments on this measure?

(Larry Glance): So again, we don't – look, we're not rating reliability, just rerating validity, right?

Karen Johnson: Correct.

(Larry Glance): OK.

Karen Johnson: So if you guys will go ahead and do that now. OK. This is our sixth measure. So we're actually doing really well. We have about 18 minutes, so let's see if we can get through. Again, if we don't, that's OK. We can come back in our next call.

This measure is also a new measure and, like the previous two, looking at kind of a special population of people. This measure is looking at basically

transition SNF community among long-term SNF residents who are enrolled in Medicaid or managed long-term care plans.

So that – again, with this measure, I think there was a lot of confusion about whether or not there are exclusions to the measure. Again, that's – sometimes (this one is in) the eyes of the beholder, but they tried to spell out what they counted as a long-term SNF resident that there's a number of stay – number of days that are included there.

They talked about what they would count in terms of people who were newly admitted versus people who had already been admitted, that sort of thing and people who had no opportunity for discharge. They said those are not part of the denominator.

So the question there I think is, you know, maybe there needs to be some more clarity in terms of exclusions. Again, the way they set it up they don't – all of those people that they are not following into their defined denominator they're not seeing those as exclusions, sort of saying they're not part of our denominator.

So in terms of reliability, this – I want to make sure I'm very clear here – this is a health plan measure and there – the ratings for reliability and validity were split. So we have to talk about both of these and we're going to have to vote on both of these.

So there were two moderate and two lows on reliability. Again, this – the exclusions, this was probably even more confusing because there were so many kind of details about who is included in the measure. They did do score-level reliability testing.

In terms of validity, again, a split vote. They only had – it was basically four health plans that, among those four health plans, it was 10 (minds). I forget the exact terminology that they used. It helps – it helps plan product lines. They had 10 product lines that they were able to include in their testing.

So let's start out with reliability and, specifically, the question about exclusions and what you might want to see there if there is anything or if you

feel like there is – that's something really that we need to talk about. And then, in terms of the reliability testing, they did do the signal to noise.

They quoted Morris method or Morris paper by Morris. That was a little new to me. I don't know if – how different that is from other ways of doing signal-to-noise analysis.

But their main reliability for plans that had more than 10 enrollees was 0.52. The range was pretty wide from 0.34 to 0.85 and, for the most part, that seems to track with sample size. So there was differing opinions on basically those results, so (there was too) low or not.

And then, another statement that we wanted to point out, the developers said, quote, "These measures are expected to be used for external – or internal quality improvement purposes, not payments." So we just wanted to remind you that measures that are endorsed by NQF, by definition of being endorsed, we believe they are suitable for internal QI, as well as various types of accountability programs.

So – and accountability programs that are – it's not just payment. There are lots of other types of accountability programs, verifications, networking inclusion/exclusion, reporting, those kinds of things.

So I think this – as you're thinking through here if the measure actually seems suitable only for internal QI purposes, then it really shouldn't be endorsed. It doesn't mean that it would have to be used for payment, but it could be used for payment.

So let me stop there and let you talk about reliability and the exclusion question first, and then we'll move to validity.

(Larry Glance): So I can start out the discussion, Karen.

Karen Johnson: OK.

(Larry Glance): So I'm going to approach this from – we're going to talk about reliability first, but I'm going to approach this from a slightly different angle. So the risk

adjustment model was extraordinarily limited. It only had a handful of co-variants, only two co-morbid conditions.

And so, when you're evaluating reliability, if you don't have adequate risk adjustment, essentially what you're doing is you're looking at the measure reliability almost as a – an un-adjusted rate, which will inflate your observed reliability. So despite the fact that the reliability as measured by the signal-to-noise ratio seems acceptable, I would challenge that based on the fact that I think that the risk adjustment model is probably inadequate.

When I looked at the model, like I said, it – the final model had age, sex, dual eligibility status, prior hospital utilization and two co-morbid conditions. And it would seem to me that this kind of a patient population is probably pretty sick and that probably does not qualify as adequate risk adjustment.

And then, when you look at the way they validated their risk adjustment model, they only used 550 patients in that – in that validation dataset which really seems wholly inadequate again to evaluate a model performance.

So I know this is not the usual thing, but I actually would rate their reliability lower based on the fact that the risk adjustment is probably not adequate.

Karen Johnson: Thanks, (Larry). And that's fair and we could go ahead – and to open that door, we could go ahead and talk about validity at the same time. They did do some construct validation and (bound things) going in the direction that they hypothesized.

But also, it's kind of the things that you've already mentioned in terms of the risk adjustment and the small sample size, questionable (hosh a) bunch of statistic, and then, of course, the incorrect interpretation of NQF's position regarding social risk factor consideration and risk adjustment model. So all these things are in play as well.

And I don't – I'm not an expert by any means on managed function care plan. I don't know how many there are. I don't know – I think they used to be a fairly rare breed, but maybe they're not so much anymore.

Female: Definitely growing.

Karen Johnson: OK.

(Christie Splend): I mean, I think it's concerning. I agree with the previous comment that, especially with this population and releasing to the community, it's really critical to look at their activities of daily living and functioning, their ability to walk independently, go to the bathroom independently. I mean, those are critical factors that would affect whether somebody was released or not and that's certainly not captured by age.

There are also a bunch of frailty conditions that are typically adjusted for, for these types of patients – frail elderly patients. So, yes, the model itself is concerning.

Karen Johnson: Probably more exclusions as well.

(Jean Nuccio): Yes. This is (Jean). I guess I was a little – a little bothered by the statement on (S8) denominator exclusions, none. And then, when you get down to the description of exclusion denominator details, it says "Do not include, do not include, do not include," which sound like exclusions to me. But, you know, just maybe it's just wording. So I was concerned about the consistency of the – of the model.

Also, I was having a little trouble with their concept of measurement year. Is this a rolling 12-month period or is this only annual or is it some sort of consistent semi-annual like from July 1 to July 30? And I was – I just – I was looking for more detail there.

Karen Johnson: And I don't know if anybody has an answer to that. I was just looking at – yes.

(Jean Nuccio): So, I mean, in one case they are talking about October 30 of the measurement year and then later on with the denominator they talk about – I think they talk about June something of the measurement year.

Karen Johnson: I remember that you pointed that out, (Jean). Yes. And it is – it is hard to tell if that was just a typo and they just typed in the wrong month.

(Karen, Mary Beth), do you have any thoughts on this one? (Christie)? So, (Christie), you've already mentioned yours I think.

(Christie Splend): Yes.

Karen Johnson: (Day deals) and frailty.

(Mary Beth Varcor): So this is (Mary Beth). I agree with what (Jean) had to say about the exclusions. I think they need to look a little – at least understand that sometimes you just won't get those detailed calculations. You'll just get the exclusions (and the next day none) you're not – you're – somebody is not going to pull on the mouth, yes.

Karen Johnson: OK. Any other things you'd like to discuss? (Is) the – just to look back on the ratings for validity, somebody rated it high and then there was one moderate and two lows, so it really was kind of across the board. Does the person who rated it high want to make a case for why you're OK with the risk adjustment approach?

(Mary Beth Varcor): This is (Mary Beth). I think that was me, but I think I (want to impede) sanity at the – at the time because I was just (doing search and see).

Karen Johnson: That's why we're having these calls. I think it's impossible sometimes to see everything and every little detail, so it's great to kind of discuss amongst ourselves I think. Hopefully you guys agree and find this useful.

All right. I'm not going to belabor this. I think we have hit the major points for this one. We would like you to vote both on reliability and on validity.

OK. And, with that, we have our other seven measures that we did not pull for discussion. Miranda gave you a chance at the beginning of the call to pull them and you didn't. But now that we've had these conversations today, do you – does anybody want to change your mind? Do you want to pull any of these for a discussion and potential revote?

(Larry Glance): Were these measures all endorsed, Karen? Yes, there's …

Karen Johnson: The other ones. Let's see. Let me just go through. Give me just a second.

Female: They're all new.

Karen Johnson: (Inaudible) measures – they're all new? OK. Yes. They are all new. Let me make sure that's completely correct. And I'm saying that now because I think (inaudible) was wrong.

There was – there was one measure in one of our (inaudible) forget which one that it looked like it was a new one, but it actually had to do more with our numbering because it's somewhat a little bit kind of wonky there and it really wasn't a new measure. But all of these really are new measures.

Female: I think (the) question (inaudible) we vote up or down on this.

Female: They don't pass the current voting.

(Larry Glance): Yes. OK.

Karen Johnson: Yes. So if you don't …

(Inaudible)

(Larry Glance): … yes.

Karen Johnson: Yes. If you don't want to pull them for any particular reason, then we would put the votes through as you see them on the discussion guide.

(Larry Glance): OK.

(Karen Troymatic): So the only other one (inaudible) sort of – isn't there one – just like the one we just discussed?

Karen Johnson: There is one pretty close to it. It is the – I think it's the hospitalization for ambulatory care sensitive conditions for duals. Is that the one you're thinking of?

(Karen Troymatic): Isn't there another admission to an institution – admission to an institution. That was – (it was a flip), but it's the same – it was the same presumably developer and the same testing with the health plan product lines.

Karen Johnson: Yes. 3456, that first one. And then, I think probably the same group minimizing the length of stay. Did you want to …

(Inaudible)

(Karen Troymatic): … has the same – because I was just curious is I voted moderate on them, other one although I was swayed by the concerns that risk adjustment (inaudible) (great). But is that not a concern for those other ones? I just (feel) we should be consistent.

So if they're basically the same measure, is it worth being sure that we're being consistent across those three unless there is something different that (met) – the – obviously, one of the criteria that we've decided on is different. I just don't know without looking back at them (to tell you that).

Karen Johnson: Yes.

(Larry Glance): This measure was risk adjusted I think, Karen.

Karen Johnson: Yes. So I'm looking at 3456. It was risk adjusted via stratification, so they split it out in four age groups. So a little bit different approach, but really only age brought in there.

And it looks like people mostly were concerned about – one of the things that came out with the low event rate, so we have a bit of a (wear) event potentially. Is that right? I don't know – I don't know if that's correct or not – concerns about the risk adjustment, small sample size, the – and then, kind of a more overarching concern about what the measure actually does.

Does this really reflect quality? And I think it's something that we would want to, if this measure does go forward – it's something that we would want the standing committee to discuss.

So what do you think, (Karen), do you – do you guys want to pull this? We'll come together on a call whenever our next call is scheduled and we'll discuss – look – it'll give you guys time to look a little bit more closely at these?

(Karen Troymatic): Maybe I'll do it by e-mail. I just wanted to be consistent. It just occurred to me that if they have the same specifications we should be consistent. If there are different specifications, then I think the votes are fine. I just didn't – I couldn't quickly enough figure that out, but I bring that up for consideration.

Karen Johnson: OK. Yes. That one and 3457 are both looking at Medicaid managed long-term care (and all this). Yes. So why don't – why don't you guys take a look at 3456, 3457; take a look at the risk adjustment approach for both of them and see where you land?

In the meantime, we should be able to know what your votes are on the other ones. (So we all) officially know what you guys have decided on that other one yet. And we might be able to – if we can resolve through e-mail, we will. Otherwise, we'll come back on the next call. Is that good for everybody?

Male: That's good.

Female: Sounds good.

(Larry Glance): So you will notify us if we have a next group two call?

Karen Johnson: Yes. So what …

(Larry Glance): OK.

Karen Johnson: … we'll do is we'll look at the writings for the other measure that we just talked about because the concern really is are you being inconsistent in your application now that you've talked about that first measure. We'll let you know kind of where the votes landed.

You can go back and look especially at 3456, 3457 see if you do think they are kind of the same concerns and we can decide from there if we need to actually officially talk about them or not. While you're at it, I would suggest

go ahead and just take a peek at the rest of the ones on that list just in case something has occurred to you that you feel like you do want to talk about it.

(Larry Glance): Karen?

Karen Johnson: Yes.

(Larry Glance): So I'm going to be in the OR for the next – the next conference call that you scheduled, but very quickly, looking at my notes, for 3456 – so one of the main problems I had with the other measure was that the risk adjustment I thought was problematic. 3456 was not risk adjusted. It was a stratified measure.

And then, with 3457, I had much less concerns with this particular model. They – instead of including just two comorbidities, they included a wider range of comorbidities. They had a bigger validation dataset. Instead of 550, they had 3,400 patients and they did look at the C statistic and – which was, although low, 0.63, is not completely out of the range of what one would see with these types of models.

And they also looked, I believe, at – (I'd say) I don't know that they really looked at calibration very well. They looked at the HL statistic, but again, because it's such a small sample size, it's not a very good way of looking at calibration. I just wanted to make those comments quickly because I don't think I will be on the next call.

Karen Johnson: OK.

(Larry Glance): So I think that these two measures are different enough from 3456 that we could vote on them differently because they're not the same.

Karen Johnson: Thank you for that, (Larry). That's great and that will help people as they go back and check real quick. So we'll send you an e-mail with those results. You could take a peek and we'll decide hopefully pretty quickly. If we do need the call, we'll keep it. If we don't, we'll (inaudible) and give you guys two hours of your day back.

So I know we're a few minutes over. Thank you so much for joining us today and doing all this work. We know it's been a very heavy lift.

If you have feedback for us about anything that we've done so far in terms of our new process, subgroup discussion, (getting) to talk about things as a group, voting as a group, how the process actually work, anything like that, any kind of feedback you'd like to provide, we're more than happy to accept it.

If – does anybody have anything else you'd like to say?

(Larry Glance): I thought that went surprisingly well for this amount of measures that we had to evaluate.

Karen Johnson: I didn't think we'd get through them. I really didn't.

(Larry Glance): I didn't either. I really didn't. I was – I was sort of pessimistic about this, but it worked.

Karen Johnson: Yes.

Female: I will say I think having the opportunity to discuss them as a group is really helpful. I feel like I learned something from this in addition to feeling more comfortable about my ratings for each one.

Male: Yes.

Female: And I agree.

(Larry Glance): I agree.

(Jean Nuccio): And did the first group do well also or …

Karen Johnson: They did. We are coming back with one more measure. Somebody actually pulled one at the last minute, so we've got through the first ever how many we got, then we pulled four that we pulled initially and then we pulled (a bit more) at the last minute.

But it's going to be something that I think that will be really useful to everybody kind of across the board. So we're glad the issue came out and we can kind of (thrash) it out, but they did well too.

(Jean Nuccio): And we will – we will discuss the sort of consensus within the consensuses – (consensi)?

Karen Johnson: Yes.

(Jean Nuccio): It's been a long time since Latin.

Karen Johnson: Yes. I was actually just thinking about that before our call and what we need to make sure that we provide back to you. I mean, at minimum, we'll provide the overall ratings and the summaries that we come up with. But I think there might be – maybe some other things that we can provide to you.

Like I'm taking notes. Obviously, it – kind of the way where we started on this call today, there is a couple of things that have come up in the first subgroup, in this subgroup, that I know are going to come up in the other subgroups that are going to be kind of (fotter), if you will, for some of our monthly calls, the things that we need to do, so I'm taking notes on those.

The first subgroup really pointed out a couple of things that they weren't happy about in terms of NQF's requirements. So thinking that maybe we are a little too easy, to be honest with you, and would like us to make things a little – to add some more requirements. So those kinds of things also we'll be talking to you about.

OK. Well, thank you so much. We will be in touch with you regarding those two measures and …

Female: Please submit your votes if you have not already done so.

Karen Johnson: Yes. I bet you everybody has already done their votes, so thank you all so much.

Male: OK.

Female: Thanks.

Female: Bye.

Female: Bye.

(Larry Glance): Thanks, bye-bye.

END