National Quality Forum

Moderator: Scientific Methods Panel
October 09, 2018
2:00 p.m. ET

OPERATOR: This is Conference #6386284

(Tonya): Thank you. Welcome everyone to the Scientific Methods Panel Subgroup 1 call. This is our first and hopefully main call. We do have a follow up schedule, but we're hoping to get everything done today.

This is currently (fleet involved) and Senior Project Manager on the project. I wanted to start out today with just some facts on how we're going to be running the call and what we're hoping to achieve today.

So first you – the subgroup members did receive what we're calling a quote-un-quote discussion guide. This is mostly just staff notes about the measure. Trying to give you an idea of what issues that came up that we think is important to bring to your attention for measures that pass through.

We just wanted to acknowledge some things that were stated by the panel but didn't really change the decision overall in passing or nor passing. So the way we have it set up is we have the (could have dot) reach measures, which are predetermined to be discussed today.

These are measures where either the panel amongst them self were not able to come to coconscious. Some voted positive, some voted negative. And we needed to have a discussion to see which way we wanted to move forward with those measures.

Some of them may have had a positive review, or a negative review but we thought as staff it's important to pull the measure for your discussion. Maybe based on (end criteria) or so on. And after that we have what we have the consent calendar of passing measures. If we do have measures in the future that are do not pass. We will have that consent counter as well.

But this one basically states that these measures are voted in a positive manner. We have a selection for them. And we do not intend to discuss them, unless a panel member does choose to pull the measure for discussion.

So that is the consent calendar for the end of discussion guide. In the e-mail we sent to you last Thursday evening it did contain an SurveryMonkey link. We ask that you pull up that e-mail with the link. Open the link for each time we're going through a measure. And put your votes in as we go through.

Just that way we can get your votes in as quickly as possible. And we can have a decision made and know if we need to meet again for our next follow up, if we need more information from you and so on.

So if you could bring up the SurveryMonkey's that would be great as well. Also just so you know we have four measure slated to be reviewed. We did not receive any e-mail from subgroup members asking the other three measures to be pulled.

We're aiming to spend about 25 measures on – I'm sorry 25 minutes on each measure. And we're going to try to keep you on track for that. We do want to get through all the measures efficiently as we can.

We do ask that when you speak up to say your name, since we don't have webinar system we wont be able to use a – like a held up hand or anything like that. So we just ask that you try to speak up when you can. But say your name so we can know who (inaudible).

And then also just so you're aware this call is public. So developers and members of the public can be listening in to the discussion that you are about to have. However there is no public commenting, nor can we open up the lines to ask a developer a direct question or anything of that sort.

The goal of today's meeting is really just to have a discussion amongst the panel, the subgroup of the methods panel. To get a decision made. Were there any questions before I give it over to (Karen) to start our first measure?

(Jen): This is (Jen) will we follow the measure in your – the order in your outline? I just want to make sure I grab, you have nurse working index first, and so on?

(Tonya): Yes, that's correct.

(Jen): OK, great.

(Tonya): We will (bring in the quarter). Oh, wait which one did you is first on your list?

(Jen): A practice environment scale.

Male: (Inaudible)

(Tonya): That's the consent calendar one.

Male: Oh, that's on the…

(Jen): Oh, you want me to go with the document? I'm sorry.

Male: Yes, this environment, diabetes, and the CMS measures are all on the consent calendar.

(Jen): Yes, right.

Female: So those won't, we won't – the nursing one, the diabetes, and the total (inaudible), those are actually on our consent calendar to be passed. So we won't discuss them unless you pull them to be discussed.

(Jen): OK.

Female: You should be able – the discussion guide as really your layout of how we're going to be going through the meeting. The first measure up is (early 753).

(Jen): Got it. I’m with you. Thank you. Sorry about that.

(Sherri): This is (Sherri), can we just really quickly discuss the nursing measure ICC? Because the way – and I was the one that said insufficient data. But the way that ICC is calculated I believe is incorrect.

And either you can actually to take it offline or what ever, but they didn’t include the patient error term in the denominator. And for multi item measures that’s really important. And it got left out.

And I’m not sure why it was considered OK? But those calculations can’t exceed the (com box) alpha for the reliability. And they did not include, at least I didn’t find it the patient level error term in the denominator.

(Jack): So this is the nurse environment scale, (Sherri).

(Sherri): Yes.

(Jack): Yes, there is no patient matter. This is a survey of nurses.

(Sherri): Sorry, I meant the survey the individual. So the (inaudible) who ever did the multiple measure, that error term should be in the denominator and I couldn’t find it.

(Jack): Yes, and I was reading the – and the issue to me was whether the nurses in units had a high degree of conscious about the scores, which is the way I thought they had run the test. This is (Jack) buy the way.   So, the only relevant ICC here is not the between units measures but the within unit conscious.

(Karen): So this is (Karen), at NQS. So (Sherri) what we’re going to do is were going to call this pulled by you. And we’re going to ask you to hold your discussion and we’ll come back to it. So we are going to pull (3450) for discussion a little bit later. Either on this call if we get to it or on our subsequent call if need to.

(Sherri): Great, thank you.

(Karen): Yes. There's still two more. The optimal diabetes care measures and the cost, the HIPPA need cost measure. Did anybody want to pull either of those two? So we had asked you to pull them before. But it's OK to pull them now if you want to.

(Ron): This is (Ron). I might just say it was a pleasure to see the – a well written response by CMS to our questions.

(Karen): Oh, great. Thanks, (Ron). All right, so it sounds like a (inaudible) help me with process here. It sounds like they OK with not pulling (0729) and (3474).

So we're going to put these two measures forward with ratings as on the discussion guide. We don't need to do any additional voting or anything like that. We've got verbal consent that this is OK.

Female: Yes, that's correct.

(Karen): All right, great. All right, so we're two measures down folks. That's great. The – are you ready to turn over to me to go to the next thing, or any more?

(Tonya): Yes, we've still got to do a roll call real quick.

(Karen): OK.

(Tonya): And see who we have. I know we already heard (Ron), (Jack), and (Sherri). (Jen) are you on?

(Jen): Yes, this is (Jen).

(Tonya): (Susan White)?

(Susan White): Yes, I'm on.

(Tonya): And then (Paul Kowalinski). Thank you, go ahead (Karen).

(Karen): OK, great. And let us know, I think probably this first subgroup call maybe a little clunky on our part. We tried to do our best to make this flow in the best way possible. Hopefully we mostly got it. But if there are things that we can do better, and you want to let us know after the call we're happy to take suggestions.

The other thing just to kind keep in mind, we know you have a discussion guide. In that discussion guide we try to pull out the main things. It's certainly not every little thing that everybody noted on your preliminary analysis. So as we go through, if there's something that we didn't flag.

But you feel like you want to bring that up. Feel free to do that. That's certainly fine to do. And we also know that you maybe needing at some point to kind of bounce back and forth between some of the other documents.

So for example you may want to have your preliminary analysis in front of you just to remind yourself of what you had put in. You may want to bring up the submission material, either the measure information form or the testing attachment. You may need to refer back to those.

And we'll be doing the same thing here. So there might be a little bit of musical chairs going on if we need to kind of go back and forth. And apologies that we don't have a webinar app form that we can do that for you. So we'll just do that best we can on that.

How we'll try to do this is different ones among us will kind of walk through the different measures. So it'll be us I think U.S. staff facilitating your conversation. We didn't require that the co-chairs attend all eight calls. So we're going to be doing the facilitation role today.

We're going to go through just real quickly some of the measure high lights. Now those will be on your paper but sometimes it just helps to kind of get your head in the game when you're talking about the measures.

So we'll do a few things there. And it's the items to be discussed that we really want to concentrate on. So in some cases we'll tell you some stuff in the discussion guide about ratings for one criteria or another.

But we won't necessarily need to discuss it unless there's something that you flag that you really want to. So we'll try to go after I do the measure high lights, well try to actually have the discussion in the order of the items to be discussed section.

To the extent that we can and sometimes it's really hard to not talk about validity if we're talking about reliability and vice a versa. It's the extent that we can kind of group our discussion. I think that would be helpful as well.

The other thing (Tonya) mentioned the SurveyMoney, hopefully you guys will be able to follow along and vote. That way you just, when we're done with the discussion on the call, hopefully you'll be done with that measure.

But that's not a live vote so we're not going to be able to see what you're voting right now. So it's not a live vote. Which means that even if ultimately for example a measure fails on reliability, if we were in the regular standing committee evaluation meeting, if something fails on reliability we would just stop the discussion there.

We can not do that. So what ever is on our list to discuss today we will have to have the full discussion realizing that your votes ultimately might have stopped the discussion earlier if we had been in a different venue, if you will.

So hopefully that's clear. I don't want to belabor any of that. So does anybody have any questions before we jump in to (0753)? OK. Let's jump into it and see where we get. So this is a maintenance measure. So it's up for endorsement again.

It is a SSI, Surgical Site Inspection measure, looking at two different types of surgeries, a colon surgery and abdominal hysterectomies. So there's really two measures included under this one measure number.   The level of analysis in for this measure is the facility as well as the state.

So both levels of analyses are there. The infections the – we're actually counting a 30 day post op period. And to be included in the measure, a facility has to have at least one predicted event from the risk adjustment model.

So you see there the minimum precision criteria as listed that are under measure highlights and I'm repeating some of the data that the developer provided for us.

And what you see here is about 2,000 out of the 3,300 facilities met the (MPC) for the colon surgery, but only 787 facilities out of 3,200 met it for the hysterectomy. So that tells us that it's a fairly rare event.

The measure does allow sampling, and the sample size isn't put out. There's not a definitive sample size, but the size must achieve, quote unquote, "acceptable level of reliability," which they have defined as 0.4.

So when it comes to the ratings among the subgroup, you guys the subgroup, for reliability three people landed on moderate and two landed on insufficient. Because there was that split that's why we have – we are discussing this measure and this criterion on the call.

And there was score level testing for the facility level of analysis and there was some data element testing that would account for the state level of analysis.

In terms of rating for validity it was four moderate, one high. So we would put forward this measure unless you guys want to vote otherwise. We would put forward this measure as passing with a moderate rating, but we are going to ask you to discuss it with us briefly on the call mainly because we felt there was a little bit of lack of clarity on the data elements that were tested.

There was some questions about risk adjustment, and we also want to pull out that – point out that with data element testing only, the highest that we could rate this measure for validity would be moderate.

OK, so with that, it's kind of a backdrop to this measure. First of all, there were some concerns about the exclusions to the measure, and specifically a statement denominator data are excluded from the SSI measure due to various reasons related to data quality, data outlier, and data errors.

Now, exclusions come up in a couple of different ways with (NQF) criteria. We could talk about them as part of the specifications. So if we talk about them under reliability and specifications, it would have to do with are the specifications complete and precise. So do people really understand what the exclusions are.

But they could also come up under validity if the exclusions themselves somehow or another potentially invalidate the measure. So I think in this case the exclusion question could come up honestly under either one of the criteria or liability or validity.

So with that, I'm going to go ahead and just talk briefly about reliability testing and then I'll open it up for discussion. The – as I said, this measure was tested at the score level for the facility level of analysis, and they used a (GLM) approach and discussed between – compared to the total variance.

And for the colon surgeries, their estimate was 50.1 percent, for hysterectomies is a 52.9 percent. You'll notice under the colon surgery, they actually did leave out a little bit of information there in terms of how many facilities met the MPC, the minimum precision criteria.

So we don't really know the exact number of facilities that had reliability greater than 0.4. A little bit later on they said that about a third of them had reliability below the 40 percent threshold, so that tells us that two-thirds had the reliability above 0.4.

For the hysterectomies, we are told not only the – I guess the – I'm assuming it's the average liability. I'm not quite sure, but we know that 83 percent of the facilities that were included in the measure had a reliability greater than 0.4.

There was also in terms of being sure that there was testing at all the level of analysis that were selected, those were done as validity testing at the state level, so we're OK there.

So really the question I think hinges on the values really of the reliability estimate. So you see that the 50 percent and the 53 percent – I'm rounding up a little bit – for the two surgery types.

So let me stop there. I think that's where the disagreement was. I think probably everybody understood the methodology and some people thought that those numbers were reasonable to push forward. Others maybe didn't.

(Ron): (Karen), this is (Ron). So I agree that was one of my considerations also besides some structural elements about how the survey was actually completed. But they did report anyway the 50 and 40 – or 53.

So in credit for that, there wouldn't be anything – I mean, unfortunately the choices are moderate or low, and that's the only thing that I really equilibrated about and ended up calling it moderate. It certainly is not higher than moderate, but I'd like to hear from the two people that were insufficient if we could as to why it was felt to be insufficient. Hello?

(Jack): Yes, this is (Jack).

(Ron): Yes.

(Jack): I'm happy to start. I was leaving some space for the other insufficient person to beep in. So here's my perspective on these measures, particularly with rare events and where the intent is for the facility to asses how well they're doing relative to others, which is you want to know with these rare events that at the institutional level that the results – when I think about reliability, it means the results are reasonably stable.

I get about the same score that my ranking would stay about the same. I wouldn't be leaping for above average to within the average range.

And I didn't see any formal testing of that or any presentation of stuff on that in this either for the performance for the risk adjustment model in terms of moving people around or in terms of the scores themselves in terms of a Monte Carlo type assessment of resampling within institutions and determining where their score came and where their rank came.

So I didn't see the data that would let me assess whether these rankings were stable, and for rare events, I think those are important.

(Ron): Yes, point well-taken. I wrote in the summary the submission presented a sparse analysis for reliability and validity, mostly relying on the fact that it was based on a widely used and valuable measure.

It would have been nice to have seen a much more detailed review of the methodology and I think you said – you said exactly that. I think this and perhaps two more we're going to talk about really did rely heavily on their experience and widespread use of the measures.

And the reason I said I was so happy to see the CMS measure earlier because they could have chosen a similar path, but they did not. They gave extensive detail to where when you got done I was just in enamored with the response.

This one I did not leave with that feeling that I felt that I knew everything I would have like to have known to give it a rating. Nonetheless with the numbers they provided, I said moderate, but there was no way I was going to go high. And I understand the rationale for insufficient, and I think it's they just didn't give us that data.

(Karen): So this is (Karen) from (NQS). Just really quickly, and (Jack), I totally get it especially for the rare events. I will make sure that everybody's on the same page in terms of our absolute requirements in (NQS). We do require either data element testing or score level testing.

And we have been and I think the methods panel has been willing to accept either the kind of signal to noise type analysis, which they did do, or the, we're going to say in air quotes, that stability type of analysis.

So we have no requirement right now that they provide that. Now, if that still makes you feel uncomfortable going forward with it, we could certainly potentially send it back and ask them to do that, or we could if you're willing, if the votes go past, but we want to see this maybe in a year or at next endorsement, something along those lines.

(Jack): Well, I think given the way these ratings, these measures are used explicitly with would you rank, are you above average, below average in the average group, I think some measure of – some way of assessing the stability of those ranking ought to be part of the testing.

And so, that's part of my reaction here. So I'd like to see that, but you signal a noise becomes a surrogate for doing that. And I've got to admit I did not see the signal to noise. I saw a discussion of optimism statistics without the optimism statistics being reported.

(Ron): Yes.

(Sherri): Let me just – this is (Sherri). I was the other insufficient person. But – so what worried me a little bit about is – following up on that is that this measure's been around and it's up for renewal, and yet they have in the data, the reporting elements section 2A2.3 (ex) of 2009 facilities.

Well, that seemed a little bit floppy to me, and then I looked down further and they didn't fill in the data. They said, quote, around one-third of the facilities that met the MPC had reliability below the threshold."

So it seems to me like the precision of those kinds of comments, really if they got – they should have the data. And so, why is it not reasonable to ask to have those data presented?

And then for the exclusions in the nominator, data quality and data errors, it felt to me like are they saying that poor recording isn't a possible indicator of quality? Are they excluding some of the lower end of town here?

So seems to me like this didn't feel like there was enough data for our measure that's been around for this amount of time for us to kind of grind through the issues that we may or may not have with it. So, I was -- I still don't think there's enough here to see for me to move beyond insufficient.

(Susan): This is (Susan). I put it moderate, but I have my evaluation open and kind of said, but at the low end of acceptability. So, I was trying to measure it against -- and (Jack), I hear what you're saying on the rankings and I totally agree.

The struggle I have with this process, is I have to work really hard to keep myself in the rules that were in play and presented to the submitter, versus where we might want the rules to go or what we might think is optimal, right?

And I don't know the answer to that, but -- so, I try to keep myself sort of -- it's -- I think it's OK. I -- (Sherri), you comment that it's been around for along time and they should do better, I think is really a good one that I really hadn't though about that much. So, thank you.

(Jen): This is (Jen). I struggled throughout the whole review cycle with what I wanted to see with versus the minimum that -- and I feel like what the group's saying here is we're crossing that line, some of this really is minimum.

It should be seen to make that decision, but there definitely were and cases where I thought, OK, I'd really like to see this, this and this, but maybe that's not fair to ask for. As you're saying, sort of within the confines of the rules of this review.

Yes, and I know these ACS measures are sitting on top of an unusual sampling frame. I think the hospitals choose what they participate in and so I think it is possibly a complicated reality that they're kind of putting this measure on top of.

(Ron): So, I think -- this is (Ron) and I'm glad someone else used the term sloppy, but I was trying to be nice about it. I -- like I said, I didn't leave this analysis satisfied, but what -- the question on the table is do we (vote) based on the minimum requirements and or do we vote on what we would have liked to have seen?

And I think that's the split between moderate and insufficient. I don't know if that's what we have to decide because we're going to be having the same discussion a couple times.

Male: If somebody can point me to the signal to noise analysis, because I did not see that. I would probably feel a little bit better here. I still don't think it's the right way to do it, but I'd feel a little better if somebody could -- if I had -- somehow I didn't see the signal to noise analysis in the same way you did. So, somebody point me to where it is in the document.

(Ron): I don't think it's there. I don't think you missed it. I don't think it's there.

Male: There's no signal to noise. There's no other report. They say we passed the optimism test, but they don't report the optimism statistics. Where are the statics that demonstrate reliability of this measure?

(Karen): OK. If you -- this is (Karen) from NQF and let's make sure that I've got my terminology correct. If you were able to pull up the testing attachment form under item 282.3, they talk about their methodology that they did. So, they say -- and I'm reading from this admission.

Reliability was estimated as the between facility variance from a generalized linear mix model divided by the total variance estimated from the same model. And then they go on to say for the colon surgeries, the 50.1 percent and for the hysterectomies, they 52.9 percent that you see in your discussion guide.

Male: OK.

(Karen): So, that's -- I think that is signal to noise. That's how I called it myself. I think (Ron's) point, that is very sparse (inaudible) of what they did.

(Ron): (Inaudible) the term sparse.

Male: OK, so thank you. No, I'm not sure it's sufficient. It's awfully low for a rare event.

Female: Well also, (Karen), if you look in that same section, 282.3, it's says X of 2000 -- if they've set the minimum precision criteria at 0.4, they say X of 2009 facilities meant that -- well, it seems to me like nobody went back over the submission and filled in that number and then they say things like around one-third of facilities. Well again, it seems to me like if they got those data, you wouldn't say around, you'd say exactly how many.

And it disturbed me that this was -- somebody didn't go back through and edit this and put in the numbers that you'd need to kind of make a decision about how much -- and that probably is easy for them to clarify if they've got those data.

(Ron): Well, I dare say across the three college measures, it was almost a copy and paste. And I think you're right. I noticed the X's too and wondered, that's kind of sloppy too. Now, I mean, how do we want to handle that? We can ask clarification, then you say, we send it back for clarifications.

(Karen): I mean it's not -- so again, this is (Karen) from NQF. It's not a complete kiss of death. We now have submission cycles and evaluation cycles every six months. So, if you guys say no, we want to see some more information on, for example, the methodology that was used. We want the X's filled in there and know exactly what numbers they are.

(Jeff), to your point, even though it wouldn't be a requirement it sounds like because this is a rare even, you would feel much more comfortable being able to discuss and rate reliability if you had that some kind of stability type analysis and we might need to talk a little bit more about exactly what that might look like.

We can -- you could decide to put it forward as insufficient, we would provide that back to the developers and hopefully they would bring it back next cycle with those things clarified.

We would try to make sure that, to the extent that we could, given availability, et cetera, that you guys would get to see this one again. So, it would just be kind of checking and go from there.

Alternatively, you could say, well, OK, they did do a signal to noise, that meets our minimum requirement, the values were somewhat low, but we're willing to live with them.

And you could put forward as (inaudible). So, it's kind of up to you at this point as to what you feel comfortable doing in terms of saying that the methods panel feels that this -- you're either moderately confident or you're unable to make a rating.

(Ron): Yes. And this is one of the first discussions we had where they're part of our role is to educate and improve or to judge based on what's submitted and not raise the bar as time goes on. And I can see an argument. There is a point -- despite this meeting minimum standards, there is a point to know we're trying to do -- get better than that. I don't -- if we want to.

Female: What's the implication for the measure developer if we rejected this now, measure becomes un-NQF endorsed for six months until they try again?

(Karen): Yes. No, it wouldn't be that. I think we have some decision rules in house about how long it could push, but you -- the would certainly get at least the next cycle and maybe even another cycle to bring it back.

So, they actually -- to be honest with you, it would still look endorsed, but kind of in a kind of endorsement limbo almost. And that would be assuming that they wanted to bring it back. If they decided not to bring it back, then they would loose endorsement.

(Sherri): (Karen), this is (Sherri), can you explain our relationship to the steering committee, the readmissions committee, standing committee? Is this feedback to the developer or is this feedback to the standing committee?

(Karen): It's -- it depends on which way you go. So, if you -- if the majority of you go insufficient, then we will write this up and provide this to the developer and the developers may also be listening on the phone. We don't know. We don't have a way of checking that.

But, that detail would go to the developers. The standing committee that would be looking at this, in this case it's the Patient Safety Committee, number one, (Andrew) is staffing that, so he knows what's going on with this measure.

We would tell the Patient Safety Committee that the Methods Panel did not pass the measure on reliability and provide a very brief rational.

So, we wouldn't go into major details, but we would probably mention that the lack of stability given it's a rare event, the lack of some of the details that you wanted in terms of methodology, et cetera. So, we would provide a very brief rational of what happened.

If you guys land on moderate, then everything that we would have told the developers, we will be telling both the developers and the standing committees. We're going to be writing a summary of your analysis and discussion. So, all of that would go to the committee.

(Sherri): Thanks.

(Karen): And just so you know, and I know that it is tricky, I think it's really helped us this time, this subgroup thing, I can already see things that you guys are really wanting to see.

So, even though, right now, we have our current standards that we can't kind of play outside those too much, very soon our monthly calls we're going to go through and get your recommendations on some very, very concrete things.

So, we'll have case studies that we can refer back to and I think you guys will be able to provide some very concrete suggestions on the types of things you'd like to see. Some of those things might need governance approval if we actually need to change our criteria.

If that's the case, we would probably try to make that happen in March and it would be implemented next fall. So, that's kind of the timeline for those kinds of things depending. So, you will have, as (Ron) said, a chance to try to up the bar in certain ways coming up.

Well, I think we -- unless anybody has anything else on reliability, I think we've hit the major points on reliability. Are you guys OK with going ahead to validities?

Male: Yes. Does the discussion of the risk adjustment model fall into the reliability or validity?

(Karen): Under validity.

Male: OK.

(Karen): Yes. We see the risk adjustment approach as a potential threat to validity. So, if the risk adjustment isn't reasonable or adequate then it threatens the validity. In terms of validity, first of all, a lot of facilities didn't meet that MPC, it is a rare event.

Really, our question to you is, that is what it is. That's how they specify the measures. It's something that we need to highlight for the standing committees, I think some people call them the content committees, but that's something that we need to ask them specifically to think about as they're looking at the measure.

In terms of the testing that was done, they did do some for the states. They did do a little bit of data element testing. To us, it was a little unclear what exactly was tested.

The descriptions were state methodologies and sampling practices there in general but (ours is) reviewed (post-op which is) medical records (for) signs and symptoms and determinations made as to whether the patient met criteria for the (NHF FM FFI).

And then they compared it to what was actually included in that registry and (computed) the measure. When I first read that, I thought that that was basically they’re saying that they looked at the numerator and then later I wasn’t sure, so I just -- personally I wasn’t exactly sure what was being tested there.

There is somewhat low sensitivity based on the results. And one of you guys had the question about does that actually indicate some systematic underreporting of the measure and if so, is that a threat to the validity that would either concern you or that you might want to push on to the standing committee to discuss.

Going back again to (Sherri's) question about the poor quality data and the exclusion, and then finally the risk adjustment concerns, the ones that we had here -- (Jack) has already mentioned -- they did some optimism stuff.

That was new to me. But they talked about it being done but didn't actually show those data. They did not, apparently, do any testing of variables that were included in the risk adjustment model. Our criteria do say that all critical data elements should be tested if you're going to use data element testing.

So the question for you would be do you consider the variables in the risk adjustment model to be critical data elements. And if so, then we would expect testing for those.

And then finally, there was the odd result, if you will, for the Hosmer-Lemeshow statistic that could reflect a calibration issue for the hysterectomy surgeries. So those were the things that we found. (We'll) open it up and let you guys discuss as you want. Just as a reminder, again, this measure -- and actually, it's a really key reminder -- all of you guys passed this measure on validity.

So it was four moderate, one high. So it's not that any of you guys thought that these questions and concerns were insurmountable, but we thought that they at least deserved some discussion before we pushed them through.

So I'll stop and let you guys start. And if you guys are fine with us saying hey, we saw these things but we're OK with that Hosmer-Lemeshow test being a little funky there, that's concerning but it's not a deal-breaker. Same for testing of the risk adjustment variables, those kinds of things, that's absolutely fine.

If you guys are -- basically what we're trying to get to here is is there anything you guys want to discuss, particularly that would make you maybe vote in a different way when it's time to vote, or do you feel like it's good enough, it should pass validity?

(Rod): This is (Rod). I have the same feeling that I stated before. I would not give them high. But for moderate, I'm OK because it's -- I don't know. You're right. The thing that bothered me also was -- and I voted moderate, as you can tell.

Social risk factors were not specifically included due to data entry burden (or sighted) lack of evidence that supports the hypothesis that data collection (of such) would justify inclusion. This is a surgical model. I had some doubts about that but oh well. Their expert opinion panel said no, we don't have to mess with that.

So I think it's pertinent to our reliability discussion is do you leave feeling that this is the best example of a measure you could possibly have, and I don't think so. But is it just good enough? Probably.

Female: (Sherri) -- can you clarify something for me, (Karen)? The way they described the missing data, that's what got me really confused about what was in the denominator. It says missing data is not a problem.

(Business are enacted to) prevent facilities from entering incomplete data. And yet back where it talked about what gets left out, they said incomplete data get left out. So I'm still confused about what's in the denominator of this.

And are facilities that do a suboptimal job of recording and therefore possibly not very high caliber facilities get excluded from the denominator because of these (quote) business rules that don't allow incomplete records. So we're already got a denominator that should have better quality than -- if that's true -- than people who aren't in the denominator because they have missing data.

So that's one of the issues I had. Then I was not really whelmed by the C-statistic. And then the calibration issue for the hysterectomy is worrisome, but I was not going to weigh in on that one without some external support from my colleagues.

(Susan): This is (Susan). I think that looking at the calibration statistics, they're reporting them but they're not saying if there are concerns or not. And I feel like it's checking the box versus really doing an assessment.

But I, like you, I think seeing the (decile plots) would help us a lot more than just seeing that one statistic. But I don't know that we require that in that section.

(Ron): And (Sherri), you're right -- this is (Ron) -- that's how I interpret it too, that if you're a sub-performing institution, you're invisible to this particular measure.

(Karen): This is (Karen) from NQF. It sounds like, to me, that that is something that you guys would like to highlight and have the standing committee really talk about that a little bit more.

I think there's going to be folks -- and (Andrew), help me out -- there'll be folks on the patient safety standing committee who are pretty familiar with this (NHS) and -- yea, I got the letters backwards.

(Andrew): (Yes). So hopefully they'll have some context in which to evaluate this outside the (slim) submission form. I don't know if we want to encourage. I'd prefer them to include all the information there.

Female: (Karen), can I (come up) with a procedural question, just real quick. If this group votes reliability is insufficient and validity is moderate or whatever, can approve a measure that's unreliable but valid? It seems to me that that's a real strange situation.

(Karen): Right. No, we wouldn't. So if it goes down on reliability, then the measure goes down. So even if you think the validity is OK. And what we would do is hope that they would make the additions, revisions, et cetera to the submission and bring it back and be able to (assuage) your concerns on reliability. But they have to pass both.

Female: Thanks.

(Karen): And I think, too, what we would do -- I'm not getting the flavor, at least right now, that you guys really are majorly concerned about validity beyond (Sherri's) question, particularly about the underreporting issue.

If that is truly the case, then we wouldn't even need you to vote on validity. We could let the measure go through as moderate validity. We would still note the things that you have already mentioned, in terms of things that you would like to see.

For example, (risk F.L. plots). They are not required, but it's certainly fine for you to say hey, I'd really like to see that, maybe they could bring that back next time. So what do you think?

Is this something you guys want to vote on because you feel like you would actually change your vote or are you OK with it going through as moderate and we'll have to wait and see what happens with reliability.

(Jen): this is (Jen). Just one comment. I'm not inclined to change my vote, but I would call this risk adjustment light. I would not say that this risk model is offering a whole lot. But I don't run a hospital, so I don't know how much I really have to care when it comes to surgical sight infections.

So to me, that's a substantive question, not one that I feel qualified -- you see light risk adjustment, you see much more better models. So anyway, I don't know that all surgical sight infections just should be avoided and there isn't a lot of severity that comes into play. I was thinking that might be the logic here.

(Karen): But that's a question that if it does go forward, you'd want the standing committee specifically to talk about?

(Jen): I'd want them to feel comfortable with that, if they're the type of folks that would go at risk or be (rated) on this measure.

(Karen): And let me probe just a little bit. You're thinking more along the lines of severity factors or things like that. So not so much the C-statistic or what have you for this particular model, but just actually what's included.

(Jen): Yes, it's a pretty lean model and…

(Karen): OK.

(Jen): Yes.

(Karen): OK. We can do that.

(Jack): So this is (Jack). I rated validity high on this when I was thinking -- I was not thinking about the risk adjustment issue at all. For the places that have enough volume, I assume they can count the number of surgical cases and count how many infections they had. So I didn't see any real problems in the core (raw) ratio.

The risk adjustment model is generating the denominator for actual to expected. And the risk adjustment model has a low C-stat, which sometimes they just do. But I'm just very frustrated.

The optimism of a measure -- of analysis was supposedly some indication of how stable the risk adjustment stats are based upon the estimating -- the coefficients and the risk adjustment model. And we didn't see any of that. So after a half page of discussing optimism statistics, to not then report them or even comment on them is bizarre.

Female: Apparently something stopped, like it was (missing).

(Jack): Yes. So I suspect from just -- there's a threefold -- depending upon whether you take the bottom of the confidence interval or the top of the confidence interval into your estimate, there's a threefold shift in the estimated predicted rate.

That's a large variation in prediction, and it's not clear how stable those are from the data they present, which is sort of what I want in a risk adjustment model, I wanted to be reasonably stable in predicting the individual patient level risk.

I'm willing to go with moderate validity here, but with some pointed complaints about the way they reported the results.

Female: OK, fair enough. So I think the – what we'll do is put this forward as moderate for the – for validity with all of the caveats and suggestions for additional information that you've asked for.

And again then we'll see where it lands on reliability based on your vote. So just to reiterate, we would like you to vote on reliability, but you do not need to vote on validity.

Female: And if you can submit your vote right now, that would be very helpful for us as staff so we don't have to hunt you down later. Thank you.

(Ron): Yes, so I've been waiting for this minute, this is (Ron), so I did and the sentence said you're going to have to go back and log in again, which apparently you do, because…

Female: (Inaudible) now…

(Ron): Come to a screen that says pause to help a puppy. And so yes, I've been waiting to see if that was going to happen. So we have to go back – you have to go back to the e-mail that you sent with the link.

Female: Yes.

(Ron): And you have to open up the link, yes.

Female: (Inaudible) you have to submit your thing and then you would have to go back to the e-mail to pull the links up, that's correct.

(Ron): Yes, she …

Female: Well it's telling me I've already taken the survey, so I don't know that we can get back in.

(Ron): Oh, uh-oh, I haven't gotten that far.

Female: So if you've already taken the survey, goodbye.

Female: Yes, you answered one measure, you're done.

(Ron): Wait which one's the next one, 2456?

Female: Yes.

(Ron): So if we go into 2456, no that – it allowed – I go back and log in all over again, follow the link and get to 20 – I get – I'm in the rating page now for 2456.

Female: Yes, I have – I have no log in because I don't have a SurveyMonkey account, so I don't –

(Ron): Oh I'm sorry, I say log in, I mean just follow the link and you're there.

Female: Yes, I am and it's …

Female: That's not what happens for me. It says …

Female: Me either.

(Ron): Yes. I'm ready to vote on 2456. I'm not going to.

Female: So we'll go ahead and try to figure out what's happening with that. In the mean time we'll continue discussion and hopefully we'll have an answer for you before the call ends.

Female: And if you guys, somewhere on your piece of paper would just write down what you would have voted if you could have been able to vote, that way when we get it all straightened out, hopefully it'll be a fairly simple thing. We'll work with you offline to figure it out.

(Ron): I've been wondering that question thought for 45 minutes, now we've got to test it.

Female: All right, so the next measure is somewhat of a difficult measure to – or I'll speak for myself, it was a little bit hard for me to get my head around this measure. It's a number of unintentional medication discrepancies per patient.

It is a maintenance measure, so it is endorsed, it is an outcome measure, the level of analysis is the facility level. It is not risk adjusted, sampling is allowed, the recommendation is that at least 25 patients were sampled every month or approximately one a day.

There are concerns I think just in general about some of the definitions that were used in the measure, so there might be some frame to be a little bit more precise in some of the definitions.

In terms of reliability, you guys voted – four of you voted moderate, one of you voted low. What we are doing is we're kind of – if there's an obvious tilt one way or the other, that's how we're going for our ratings.

So we are saying that we would put this one forward as passing with a moderate rating. Now again, if you want to talk about that, we certainly can. In terms of the reliability testing that was done, it was done at the data element level.

It was a small sample size, 19 patients from one hospital to test the histories that were done and to test the scoring system. It was four patients total, one from each of the four hospitals.

And then we're given some percent agreements in terms of the history, looking at about 77 percent. In terms of the scoring system, they gave us percent agreement but also gave us, as we would want, a (Kappa) which turned out be 0.64, which according to (Landis and Cook) classification, is substantial agreement.

So again, we're – we can certainly talk about reliability if you want to, otherwise just put that forward for your information. However, for validity, you guys were across the board.

Female: Hi (Jeri).

(Jeri): Yes?

(Sherri): (Jeri), I was the one who voted low and I'm concerned for this reason that the sample size, given it's a maintenance measure, the sample size was really small and the definitions were all over the place.

I should have probably said insufficient, but I don't know what a trained, quote, "study pharmacist" is. There are all these (inaudible) that weren't provided, and given that this is a maintenance measure, you would think that they – that kind of information would be really readily available. So I still have concerns about this reliability at this measure.

(Jeri): Was that your main concern that the wording et cetera, or were you also concerned – well and the sample size. So – and going back to what are out minimum requirements, in general I think NQF would love to see additional testing over time that expands, but it's not a requirement. So they …

Female: … but is this a maintenance measure?

(Jeri): It is.

Female: So over time, they've already had time, but some of these data should have accrued. So this is – the sample size being this small and not being a brand new measure, which would make sense, that this is kind of – this is – that this level of testing was done for a maintenance measure seems to me like inconsistent with NQS guidance. But if (inaudible) happy with, then I guess whatever.

(Jeri): Yes, well happy with and does it meet our minimum requirements are possibly two different things. I think that the question for you guys is given the low amount – the low sample size, the small sample size regardless to be honest with you whether it's maintenance measure or new measure, do you still feel that that's good enough for you to be able to say, yes I'm moderately confident or I'm not confident or I don't have enough information to be confident about the reliability of this measure.

So that's where we are. Scope definitely comes into it in terms of sample. So why don't we – let's definitely talk about validity first since we know we have some concerns there. We'll circle back to reliability and see if people want to vote on reliability or if you're OK with it going forward (with moderate).

So in terms of validity again, too moderate, too low, one insufficient, so therefore we have to discuss on today's call. This is one where we actually did tighten up our criteria just a little bit.

And we – because it is a maintenance measure, we no longer accept face validity – let me rephrase that. Because it is a maintenance measure, we expect to see empirical validation. Now in some cases, it might be really hard for a particular developer of a particular measure to do empirical analysis for validity.

In which case, they can still say I'm presenting my face validity to you, here's why I can't do empirical validity, and then it would be your job to decide number one if you accept that justification. If you do, then looking at their face validity results you feel that that's adequate.

(Susan White): (Jeri), this is (Susan), I was the insufficient because of the face validity just because of that reason. I thought their justification was pretty weak. And so since I – (inaudible) and let you know why I did that, but…

(Ron): And I – and this is (Ron), I was one above that to low for the same reasons.

(Jeri): OK, all right, so the two – the two non-passes are because you feel like that there's – there seems like there's something that they could have done and you just are not buying the justification.

(Jack): So this is (Jack), I actually rated this moderate because I voted roughly past the bar. The measure's only as good as the training of the pharmacist to review the records and generate the results.

And I'm prepared to – and yet the training – the training protocol or evidence of the training works is not presented and frankly I think the (substance committee) is in a much better position to evaluate that than I am.

The cap on the (inter radar) reliability while it sort of meets the acceptable standard is awfully low and I would want to see – and to me, that's a – that's a – that signals something about the quality of the training.

So I'd like to see the cap going up over time in this measure, because we don't have a trend here. And that to me would be part of the testing for validation that would (then) make me happier.

(Ron) and (Steve) (inaudible) can certainly convince that this – that I ought to downgrade this on validity, given those concerns.

(Sherri): This is (Sherri), I was the other low (inaudible) sample size and 50 percent exclusions I was concerned about as well. And then also the – just the idea that you can – you conflate reliability and validity, they did go back and forth and they didn't seem to get which is which.

Also raised concerns for me about the overall accuracy of their test results. So they repeatedly referred to the, quote, "study pharmacist". So I thought maybe this is a research project that's being reported about and intervention sites.

You know and I thought well I wonder if somebody lifted this from a – some kind of grant proposal or paper or something. It sounded to me like this is the research project, not a report about a measure in the field that's been around for a while that's returning for maintenance assessment.

(Ron): Yes, I – this is (Ron), I agree. I wondered the same question. It might not even be a good research project, by the way, but I don't know that this is in any program whatsoever, and I think the small sample size, you're right, is this is being utilized.

For some reason it sought endorsement, but I think it's being utilized very locally, somewhere in the partner system probably. But this one is at best low on validity and suspect on reliability and has a lot of room for improvement.

(Karen): so this is (Karen) again, just a couple of things. It is a little tricky. Again it – (Andrew) and I spent time when the measure first came in just trying to understand the measure out.

And we think we were able mostly to wrap our heads around it. They did provide a little bit of discussion about their gold standard pharmacist and the training exedra. That those folks go through. They actually get I believe certified am I remembering correctly (Andrew)?

(Andrew): They do.

(Karen): They included that in the exclusion section, so kind of an odd place to put it, maybe a little out of place. I don't know if that would have changed your mind about some of the concerns you have or not.

But I did want to point that out that, that some of that is in there. We pulled out a couple things that they hinted at, a systematic review.

They talked about market's two data and Leapfrog data, which makes it sound like maybe other are using this. A little unclear but it does make it sound like there might be some room for the circle.

Male: Leapfrog just, I remember actually (Missy Danforth) from Leapfrog saying they are using this measure.

(Karen): They are using this measure?

Male: To capacity.

(Karen): So if they're using that measure then there might be some kind of opportunity to do something empirical to your point about you didn't buy the justification.

The only other thing the IRR testing results, that's certainly not what we would expect for data element validation. But this was such an odd measure in terms of the two pharmacists, they compared their histories and then compared the detections.

Normally that kind of thing wouldn't, kind of check our boxes for what we're looking for, for data element validity. But we did want to open up the possibility, if you guys' thought that that was kind of enough, and that's completely up to you. Sounds like (Sherri's) saying no.

(Sherri): Well, (Karen) can you go back to point number 12? It said please describe any concerns that you have about measure exclusions. They excluded 199 of 300. So for the small sample size of this review, it really worries me that gee wiz if this is a maintenance measure and they're dealing with it this way.

And you've got a sample size that's basically half. You've excluded from population of potentials. Is NQS not – is that not a worry on your team that this is really pretty small? And then they've excluded half the sample?

(Karen): Are you getting that from the testing attachment (Sherri)?

(Sherri): Yes.

(Karen): OK, I'm having trouble locating that. Do you have it?

(Andrew): I'm trying to pull it up.

(Karen): (Andrew's) trying to pull it up.

Female: What page are you on (John)?

(Susan): This is (Susan) it's under 2B2.2.

(Karen): Oh, I'm in the wrong place, OK.

(Susan): Yes. So, when I looked at it I made a comment that there's definitely a non response by us where those that we didn't respond. Or where we don't have data, we might have more medications and more chance of a missing medication.

So, I do think that can introduce a data. I put it under my describe any concerns you have for missing data.

(Karen): And apologies, I’m still having a hard time brining it up. (Andrew) is beating me to it. Yes, they excluded. And do we know what the exclusions were?

Male: Well they excluded people who died, obviously. They excluded people who refused. That’s probably – I don’t know how you can handle that one easily. And so those are a large – those are certainly were a large number of people though.

(Karen): So that also is a potential threat to validity?

Male: Yes.

(Karen): OK. OK.

Female: The sampling strategy they just have to take 25 cases per month, right? It doesn’t it’s not a random 25.

(Sherri): (Karen) it says on page 2B2.2, what were the statistical results from testing exclusions? And they said they compared the 180 patients that they actually included, which is more than 50 percent with 199 excluded patients. And compared with excluded subjects study patients were older, had longer lengths of stay, and had more medications at discharge.

In this study patients were required to provide informed consent so the differences between included and excluded patients may have been more pronounced if this were part of a routine, or a routine part of hospital measurement, which made me worry that it’s not.

And for a maintenance measure things like whoa. But if it’s included then in Leapfrog where – why would you put that? It’s not a routine part of hospital measurement.

But it’s been used by Leapfrog. And these circumstances, and when it’s used this way this is what you get. Again you know if that’s not a concern for a maintenance measure, OK. But it seems to me like, I get the (greeshu) when you know that what you’re including is actually biased in the direction they stayed.

And then you don't give us any rationale for -- but then when we tested it this way or we tested it in a different site we got a different result for our maintenance center.

Female: No, I think it's a valid point that that could be calling the measure -- you know it's a certain threat to validity. At minimum you're pointing it out to consider. I think you'll have to decide when you vote on validity whether that along with some of the other points that have been raised (if that enters) into move you towards slower and sufficient as opposed to moderate.

Male: Just a couple of things I wanted to -- I looked up a submission form. It does look like they ask for patients to be randomly selected; those 25.

Female: OK. I missed that. Apologies that I missed that in the discussion guide.

Male: And then the denominator exclusions; it says patients that are discharged or expired before a gold standard medication list can be obtained. So those are the patients who are being excluded.

Female: So people who died or they didn't get around to doing the …

Male: Yes, that golden standard pharmacist wasn't able to do that reconciliation by the time they left.

Female: OK.

Male: And presumably again, this can only be implemented in hospitals that have this trained and certified pharmacist. So (that's sort of) limited set of hospitals that this can actually go into effect (with).

Female: Right. I took the biased in the 180 versus the 199 to mean if you were older and sicker, you were easier to capture. Right? And they weren't consenting younger, healthier people into the study.

I don't know if that's a problem or not because those are probably complex patients who need this kind of continuity. So in some sense it seemed like they were touching the hardest cases but again, I'm not a clinician.

Female: Or could be that the older sicker ones died and therefore they didn’t count them.

Female: Well they were saying their sample was older (or) stayed longer and had …

Female: Oh, the people that actually were included.

Male: Yes, the people that were included, I think (inaudible).

Female: OK.

Female: The measure is catching.

Female: So -- but again you know -- and had more medications at discharge.

Female: Right. That’s the -- yes.

Female: So by definition you got a kind of by a sample. And then the question is what does that do to your -- you know is this measure accurate? Well for what purpose and under what circumstance and for whom and then for whom is this to be used in all hospitals or only those with a trained pharmacist?

And again, the specifications, you know that raises concerns about how well specified is this -- the circumstances under which this would be a reasonable measure to collect.

Female: So there’s a couple things going on here. A few things are -- (there’s) some terms of our criteria. I think I’ve heard that you want a little bit more decision about some of the terms that we use in the measure.

You’re not convinced about the inability to do empirical testing, so you’d like to see some more of that. Probably have a little bit more information about the training that is provided, et cetera.

And then (Sherri’s) concern’s particularly about some of the sample sizes that we used and the people who were not included in the measure. The later could be certainly something that you would want the standing committee to weigh in on as well as you guys, sounds like.

Female: And I think you could make the argument that it’s hard assess validity in the context of this study with informed consent. That does really shape. It’s hard to know what this measure would be like in the regular world.

(Sherri): The other thing is is there was some variation of the agreement statistics by frequency, dose, route, and other things that they like didn’t get a lot of detail around about that.

And then if that’s floating all over the place then those kinds of errors in medication have a greater potential in a population that has more medications, one would assume.

But they may not if the medication or if the smaller number of medications include some big league hitters like Warfarin or something that has more substantial implications when errors are made.

So I -- you know again, with the small sample size (that) you can’t really (fuse) that apart but (it) certainly would raise issues about how generalizable this is under one circumstances.

(Karen): This is (Karen) again. Just on one thing on that note (Sherri). They -- while they do -- these pharmacists do look at dose and frequency and those kinds of things. The measure itself is only about the unintentional discrepancies (this charge of admission).

So they’re counting that and they’re showing those things for a QI purposes but not for the accountability purposes. So I think that might by they didn’t go into more detail about those pieces.

OK. I think what we definitely are going to ask you guys to vote on validity for this measure. Did we cover -- I think we covered everything under validity that we had mapped out.

So we definitely want you to vote on validity. Let’s circle back real quick. The liability vote. Again, four moderate and one low. (Sherri’s) low mainly because of the really small sample size.

And so what you guys want to do with reliability, we could have you vote for reliability after hearing (Sherri's) rationale for why she voted low and just ask you guys all to vote, or we could agree not to vote and push it through as moderate. So if you guys have any preference on what to do?

OK, I think what we're going to do is go ahead and out it through as moderate since nobody's making a fuss either way, but I'll give you a second to make a fuss. We're going to push it through as moderate unless somebody really wants to vote.

OK, nonetheless, we will be sure that we are marking this discontent, if you will, with the really smaller sample size. We will ask you to vote on validity and we will be writing the summaries, et cetera, and sharing those.

Female: And in terms of the SurveyMonkey, we did update the settings so at this point, you should be able to go in and fill it in as many times as you want, preferably only one time for each measure. So please go ahead and put your votes in now, in the meantime, we'll go ahead and jump to the next measure with (Karen), (anything you want to say)?

(Karen): I think the only other thing, did we -- I didn't think of it when we did the SurveyMonkey, did we provide a free text field at all?

Female: No.

(Karen): If you guys have suggestions on anything that you would like to see that you feel like we haven't covered well enough and you want just e-mail us those, that would also be helpful, we can share those. I think we got everything.

Female: And then if only you wanted to highlight anything additionally, obviously when we're -- we may not cover everything we're talking about right now, but we do obviously have your Pas and we'll be taking all of the information into account, and your worth will still be shared with the standing committee.

So just because we didn't highlight it, doesn't mean that it won't be -- eventually get to the standing committee, if the measure passes.

(Karen): Yes. Yes, and just a reminder, we will be building a summary based on your initial analysis and today's discussion and your votes. That summary will be included, but we will also be providing your stuff from your pulmonary analysis; so if they see everything -- if the measure goes forward.

OK, we're doing actually fairly well on time, because I happen to know that the next few measures I'm pretty sure are going to really fast. So I'm going to turn that over to (Andrew) to talk us though 1716 and 1717.

(Andrew): OK, thanks, (Karen). And the issues on these two are I think, pretty much the same. These are along the lines of the first measure we talked about, 753, which was the FSI measure.

These are also part of a national healthcare network, which is CBC's sort of infection tracking and surveillance system. This measure, 1716 is tracking MRSA infections, also as in the case of the FSI's standardized infection ratio, so comparing an observed rate to addicted rate.

Also includes an adjusting ranking metric which as I understand it, is sort of a reliability adjusted ranking, sort of moves the scores to the mean for those hospitals that have low volumes of cases.

In any event, we had some concerns from our reviewers here on reliability, largely I think because the measure did not have score level reliability testing and some -- I think some of the reviewers said that it didn’t have element reliability testing either.

We wanted to clarify that for NQS purposes; we do accept data element and validity testing as basically acceptable for serving as data element and reliability testing. And they have provided some results of data element and validity testing that's similar again, to the FSI measure.

There were some states that did this analysis of comparing results to the medical record and gave us sensitivity and specificity and all of the positive and negative predicted values. And it looks like the reviewers did give a passing rating for moderate versus one low for validity.

We just wanted to kind of check with you, if you do believe that empirical data element and validity testing is adequate, then sort of for consistency sake, we might suggest that it would also pass on reliability or at least again, meet our minimum requirements for reliability.

So maybe at that, I'll sort of open it up and see if there are any questions or clarifications about that. Did that make sense, that explanation of the data element and validity testing serving at least for our purposes, as reliability testing as well?

(Sherri): Well -- this is (Sherri). I would argue that you can't be reliable -- its like, OK, so you show up for practice a half hour early, half hour late, and on average, you're on time.

So you're unreliable but on average, you're accurate. It strikes me as, if you were on a baseball team, you'd get kicked off the team by the coach pretty fast. So something has to have, I think, both -- and if you guys are comfortable with -- you don’t need to test for precision.

As for accuracy, would think that's kind of a non-starter in my world and in the measurement world. And sensitivity and specificity aren’t assessments of validity, not reliability. So I have absolutely concerns about the reliability testing and -- but if you guys -- I mean, we have to kind of be guided by what you want to do, if that's your kind of policy.

(Jen): I'm the moderate and it really was because that was rules. I had a long list of things I wanted to see. So truth in advertising. I did think that based on the rules, they had met the minimum requirements.

(Jack): Minimum requirements for which, (Jen)?

(Jen): I'm sorry, I have on the reliability.

(Jack): Reliability?

(Jen): Yes. I was moderate on validity too, but looking at my notes, I'm juggling paper all over the place here, and I had this long, long list of problems I had with the reliability, but I felt like the rules made me kind of get to moderate. So I appreciate the issues that people are pointing out, I'm happy to switch my vote.

(Ron): This is (Ron); I broke the rules like (Sherri) did. I was insufficient on reliability and moderate on validity. I didn’t give them credit for the validity.

(Jack): OK, I don't see the two as closely as related as you guys are. I see the question of sensitivity and specificity as a data element issue.

Can you get the right data? The reliability to me is more -- yes the data element component is there, but the broader issue for me on the reliability is the inter-hospital comparison and stability of those and they reported no testing on that.

(Ron): Right.

(Jack): There's no testing for reliability.

(Jen): Agreed.

(Ron): Agreed.

(Karen): So we agree that they didn’t provide score level reliability and we also agree that the testing that they did wit sensitivity specificity is data element validation. So we completely agree with you.

We do have a longstanding -- I don’t know what you want to call it -- almost a shortcut, a concession, if you will, that has been in place for many years since before I came here that basically states if testing was done, validation of the data element, then we would not require additional reliability testing. So that has been our rule, it's still in play right now.

Again, we could talk about in the next few months if that's something that we want to list to that we are actually going to expect separate reliability testing, perhaps even at the score level. Again, that's all TBD in the future, but right now, I think we do need to ask you to go with our current requirement. And then so the question before you is, you see the testing they did for the validity, is that good enough?

They didn’t test every data element, but they did test some. The results are pretty good; it looks like its a little low on -- for one (state). So let me -- do you guys want to discuss any of those things? And I definitely hear you that that concession of our makes you uncomfortable.

(Jack): No, that concession of yours needs to be revisited.

(Karen): OK...

Male: Let me be really aggressive about that. I mean, these (are used) to being to compare places. Am I higher or lower than the other guy? And that's in the realm of reliability.

Do you get the same comparative standing for folks? If you're not using it for that, then you don’t need any of the measures publically reported (at all), find your inspections, go do a root cause analysis and ignore how you're doing relative to other people.

But once you’ve got that, how are we doing relative to other people, component here, you’ve got to assess reliability at the score level, not simply at the data element level.

(Karen): So we'll be coming back to that, probably in December or January timeframe to ...

(Sherri): (Karen), this is (Sherri).

(Karen): Yes, (Sherri)?

(Sherri): I think it would help is in our reviews, if there are those kinds of issues, and you guys screen these measures before you send them on to us, if you just said to us, don’t worry about this, rather than have us go through and raise -- because we try to read them very carefully and go through these things, it's hard to find things that are missing, and if you guys have rules like that, that you just want us to -- don’t -- just ignore reliability on this measure because if you get to a whatever, a validity question, then don't worry about it, we'll flip the reviews around, or something because if you want us to review them and then you have these criteria already in place, it seems like a little bit of a waste of our time to go through those kinds of issues if there are existing policies.

And then you can take off in a different conversation, we can have discussions about whether or not you actually need to measure both reliability and validity, especially for maintenance measures.

(Karen): Yes, we can try to do that kind of thing. I think maybe we did on occasion with -- at least some of the measures that we though, I think we flagged. But I don't know that we caught everything. OK, so I think what we need to do there, do we need a revote? Do we need an official revote?

(Sherri): No, if you are content that validity is indeed moderate, than we would just make the rating for reliability moderate.

But if you are not content that validity is moderate and you would like to change that rating, then that would automatically change your reliability rating. So it's really more about do you think validity is valid and then if that's not – I'm sorry. If validity isn't …

(Karen): Adequate.

(Sherri): … adequate and then that will really determine if we need a revote or not. So is there anyone who does not think the validity is at least at a moderate rating?

(Ron): This is (Ron). I'm trying to – I'm trying to make the right argument but I can't get there. I can't get to a low validity although logically we should be able to. I (get it) for all the reasons that we said, I mean it should be easy to say no, it's low validity because of da-da-da-da.

But I just can't. I think we need to point out though somewhere in the comment section that this – the discussion we had about revisiting the automatic reliability.

Female: Yes, and we will and we will also include all those different things because your work was not wasted. These measures, they are – I would think this kind of analysis is something that they could do.

I hope they can do it and we would have probably phrase it assuming if our – if our criteria do change then you guys certainly expect this kind of precision analysis, et cetera – reliability.

(Ron): But we don't have to (put) them?

Female: No I don't think so. I think – I think we're good.

Male: And then if…

(Ron): You said 1717 is exactly the same issue.

Female: It's exactly the same issue.

Male: Very comfortable in sort of carrying over our decision to 1717?

(Ron): It's got the same things. Everybody's five moderates on validity and insufficient reliability. I'll tell you this is a dangerous trend.

Female: We're getting consistency though; that's a good thing.

(Ron): You are getting reliable, yes.

Female: All right, well thank you everyone.

(Ron): I don't know how valid we are but we're reliable.

Female: OK, so then we do not need to vote on 1717 or 1715 and (Sherri) did pull 3450 so we can talk about that. We're very – we're doing so well on time so hopefully we can get done with this measure today and we're not going to have a follow up call.

Female: If we don't quite finish this measure, then we do have our second call that we can talk about. So (Andrew) was this yours or mine? We didn't really talk too much about…

(Andrew): This one's here…

Female: This is mine. OK.

(Andrew): Oh maybe it was mine, that's right – that's right.

Female: Just one thing while (Andrew) is getting his stuff under control. I can always help you too as much as you need (Andrew). We actually do have this and this actually doesn't make any difference in terms of your discussion but I did want to point out that this is technically a new measure but it's actually not.

This is a maintenance measure that's being brought back that we actually had the developer bring it in under a new NQF number because when it was last evaluated it was evaluated as a composite measure and it doesn't – back in the day believe it or not as NQF people had lots of different definitions of what a composite measure is and there was absolutely no consistency what so ever in the way that things were typed.

So anyway, long story short, we asked them to bring this back not as a composite measure but as an instrument-based structure measure. So because it is an instrument based measure we are expecting reliability and validity of both the data elements, the instrument itself as well as the performance measure score.

The – and (Sherri) your – the reason you were pulling, let me just make sure I understand it, you – you were having difficulty with reliability and specifically you didn't believe the results of the ICC. Is that correct?

(Sherri): Well I believe the results of the ICC. I don't believe that analysis was done correctly and here's why. When you've got a between facility, different in the numerator in an interclass correlation coefficient and then the denominator has a between plus a within facility across nurses variable.

But then when you're – you also have to include in the denominator when you're doing these higher order levels of ICC you have to include the – within nurse across items in the denominator.

So what you do is you have a big – a larger denominator than you would otherwise have so if you just did a between divided by the between plus within facility variability across nurses, that's what you get when you get an ICC done that way.

But then you also need to include the error that you have even though the (chrome box off) is pretty good, the coefficients range from .71 to .96. So .71 is on the low end of group comparisons but you need to include that error term in the denominator.

You cannot have ICCs at .996 when you have precision estimates that aren't that high. So you need to include that term in the denominator when you're doing these kinds of comparisons and that doesn't appear to be what they did.

(Jack): So, (Sherri), you're saying – so, we've got multiple scales which are then aggregated into the total score. And what you're saying is, their reporting – the analysis at the score level of the individual scale level, but not taking into account the variability, and the nurse reports across the different items within the scales?

(Sherri): Correct. And so, that's an error term. So, you've got the within facility variability that's an – you would consider an error. So, that – that's the degree to which you leave a thumbprint across nurses within a facility. But then, you also need to, as you said, include the – within nurse, across items …

(Jack): Right, but …

(Sherri): … of the variable denominator.

(Jack): Oh, so the issue here is whether each of the items are measuring the same thing, or whether they're measuring slightly different things that fall on a common domain?

(Sherri): Right. So, the nurses …

(Jack): So, the staffing may or may not be adequate. The physician nurse communication may or may not be adequate. The ownership of the practice may or may not be adequate. But is there any reason to believe that those – the answers to those three things should be correlated?

(Sherri): They're very – there's variability across items in anything. There's never any perfect agreement.

(Jack): Yes, but what I'm saying is, they're aggregated up to get a score in a given area. And the factor analysis generated the subscales.

(Sherri): Right.

(Jack): But if we ignore the subscales for the moment, the quality of physician communication, the adequacy of staffing, the ownership of practice, the individual components may aggregate up to a sense of how good is the work environment here?

But the correlation across those items, which is what you're measuring with the nurse – at the nurse level, the correlation of the individual items, there may not be any inherent reason for there to be a correlation between the staffing and the communication.

(Sherri): Well, there may be (inaudible).

(Jack): They may be accurately reporting them. The nurses on the – they will agree, communication here is terrible. The staffing is adequate. And that, to me, is the level of correlation I want to see. But I'm not necessarily worried that there's a correlation between communication and staffing.

(Sherri): It's the error term of a nurse's consistency across all of those things. So, you've got multiple items measuring each one of these constructs.

(Jack): Yes. And what I'm saying is …

(Sherri): (Inaudible).

(Jack): …nurses might, legitimately, in a unit, see real differences in performance in each of those – each of the domains for the individual items that one should not, necessarily, see consistency there, but aggregating it up – but aggregating it up, you get a sense of how good the work environment is.

And implicitly, there's the tradeoff. If a place is well staffed with crappy communication, the two, sort of, are averaged out in the aggregated scale.

(Sherri): Well, I think (talking) across purposes because I did a much more – much more (per say it) kind of concern which is not the agreement across nurses within a facility which is the – that – is already represented in the denominator.

(Jack): Yes.

(Sherri): It's in nurse across these items.

(Jack): Yes, and what I'm saying is the – given the items that are in the scale, one wouldn't necessarily assume that responses to each of the items would be correlated because they're measuring different dimensions of the work environment which could be different levels of positive and negative.

(Sherri): Yes. (You've) accepted it reflects the error term within the nurse, you got me.

(Jack): Yes, but it's not an error term. The nurse, correctly and consistent with her colleagues, said communication here is terrible, and staffing is adequate. I'm more concerned whether all the nurses on the unit said staffing is adequate, or if there was a lot of variability there.

Then, whether they – the nurses were in agreement on a zero to five scale – I think it's a zero to five scale – one to five scale, about whether each of the items was positively correlated with one another.

(Sherri): I still – I think that error term is, as far as I'm concerned, already included in the denominator and what isn't is the error term that is within nurse. So, that's – that was my basic concern. And the agreement across nurses within a facility is that within facility across the nurse's term, there's already …

(Jack): Yes. And that to me is the concern I have about the reliability of the measurement, not the within nurse consistency and scoring on each item because we only see …

(Sherri): Well, but you can't get there without that error term is what I'm saying. You can't get to the – you don't know how much the magnitude of the error is. And then, you don't know how much it's conflated, supposing that one half of the nurses are all over the place with respect to their agreement, and so, with – within the survey instrument.

(Jack): OK. Well, you can get there if you treat the assessment of each individual item independent of all the others. And if you treat each scale – a subscale score for each nurse independent of the items. And if you treat the overall score of – for – that you calculate for each nurse independent of the items.

(Sherri): Again, I – you can't – for – you can't get to an agreement that's larger than the error at the nurse level. And the error at the nurse level ranges from .71 to .96. So, you have to – that coefficient has to be part of the denominator. And you can't get there without it. So, it's – these ICCs are way too high. They can't be greater than the precision at the nurse level.

(Jack): OK. Well, then we should ask them to review this and then come to us with it. I'm happy to defer that.

(Ron): This is (Ron). I'm going back. I have to admit, I did not go back to (Lake)'s original article in 2002 where this scale was developed. And I'm trying to see if the answers to (Sherri's) -- the answer to (Sherri's) question should have been provided if that's true, and whether or not it was just overlooked as far as a …

(Sherri): No, I went back to that article and looked at it. It's not there.

(Ron): It's not?

(Karen): And just to be clear -- this is (Karen) -- the (Lake) articles show the factor analysis stuff for the items. What (Sherri's) talking about is the score level reliability.

(Ron): Yes.

(Karen): Yes. And this has come up in several measures and we've talked about it.

(Ron): Yes.

(Karen): So I think it's -- my tongue is a little bit in my cheek, but it's a little bit of the battle of the (statisticians), I think. I think you guys know -- you know the point that (Sherri's) trying to get and I'm still a little bit behind on that point. So I can't help you, I can't rephrase it any at all.

I think we have a couple -- so, first of all, does everybody understand (Sherri's) point?

(Jack): Well, I want to pull up her thing and read her …

(Ron): Well, in any survey -- in any instrument-based measure, you need to know the reliability of the people filing out the instrument. And that's -- I think I brought that up, oh, six months or so ago. And I don't know how -- unless it's specifically stated, you don't even know that the development of the score is reliable if you don't know that the input isn't reliable.

And I admit, this is -- it is not stated in this measure. I said moderate, but now I'm -- or, no, I said high actually, I think. It's not referenced in this paper, that's for sure. And I got to think a second about (Sherri's) -- so yes, the score can't be as high as it is if it's not -- it -- because it can't be higher than that, so.

(Karen): So (Purna Renee), when is or next call?

(Purna Renee): Our next call is actually not for a while.

(Karen): OK.

(Purna Renee): It is October 22nd at 10:00 a.m., so about two weeks from now.

(Karen): Two weeks from now. So …

(Jack): Do -- could …

Female: … (inaudible) not here.

(Ron): …. can we get an answer to that question by then, or not?

(Karen): What question, (Ron)? I'm sorry, I lost your question.

(Ron): Whether at the nurse level they have data about the in-nurse variation.

(Karen): In-nurse and you're talking about something different than Cronbach's alpha reports …

(Ron): Yes.

(Karen): … and different than factor analysis, am I understanding you correctly?

(Ron): Yes.

Male: And …

(Ron): Correct, (Sherri)?

(Jack): … well, can I clarify with (Sherri) -- are you saying, (Sherri), that the -- they may have reported the, sort of, nurse-level reliability with that 0.71 to 0.96 but that they, then, did not incorporate that into the, sort of, larger reliability score?

(Sherri): When they wound this up, yes. It says, of the 46 articles reviewed in (Swinger) et al. published between 2010 to 2016, 37 percent -- or, 37 reported Cronbach's alpha. The coefficients range from 0.71 to 0.96 with the exception of 1.67 and 1.53 in a small sample size. They result -- the result supports the coherence of the (different) subscales.

But then, when you wind it up -- when you -- so now you need to take that into account because it's not a perfect reliability, nothing ever is in these multi-item scales, you have to -- the respondents within subject variation has to be included in the next -- in the denominator.

And you can’t, by definition, have a larger interclass correlation when you’re comparing a higher order, you’re winding it up to the next level, across nurses within a facility and then between facilities.

That extra term in the denominator means those interclass correlation coefficients can’t be larger than the bias that is at the individual respondent level.

(Jack): Yes.

(Karen): So we have a couple options in front of us, one is we could have you guys vote on reliability for this measure and kind of land where you land. And depending on where you land, we would have some feedback for the developer, potentially, of what you would like to see.

Another option that we have is to put this on hold for a couple weeks, come back in two weeks after you guys have all had a chance to think about it a little bit and that sort of thing. And just kind of start the conversation again and vote at that point.

(Ron): (Karen), there’s a third option.

(Karen): You have another option? What would that be? Put it forward (with the) …

(Ron): (Four rated) at high validity, we don’t care what reliability is. We just -- we did it -- did two of them, so.

(Jack): Yes, I’m -- I’ve now got the testing document up and I hear what (Sherri’s) saying about the table (2A2.3A), which is on page 5.

(Ron): Yes.

(Jack): And she’s almost convinced me. I’ve got to go back and think about it some more. But if you go to the table immediately below that, you see (these) in the range that she would expect -- that we would expect. So the question is can we make a decision about reliability, ignoring (3A) and just look at table (3B)?

(Ron): Yes, which does have the 50s to 60s and so on. And I raised another …

(Jack): Yes, and …

(Ron): … question, do we care?

(Jack): … Right. Well, and given the difference between table (3A) and (3B), I would be inclined to ask the developers to tell us what the hell they did in (3A) and do they want to revise that in light of (Sherri's) comments in light of (3B), and read this at this measure if (Sherri's) content with whatever they do to revise (3A) or what they've got in (3B) then we can move forward?

I think (Sherri) has hit something. The numbers are really discrepant between (3A) and (3B), and that suggests something was not done right. So we could either ignore what was not done right and make our own judgment, we could ask them for clarification, or we could vote which would lead them to provide a clarification.

(Jen): Why not ask for clarification? Why not give them a chance to answer if that's quick and easy for them to do?

(Sherri): That would be my preference.

Female: I would agree with that. We might even be able to e-mail vote on it if their answer is adequate, maybe it's wishful thinking.

(Jack): Oh, I think it'll be adequate. I'm not quite sure what they did here with the NDNQI data set, but they certainly should have access to it.

(Ron): How does that impact the four high validity ratings?

(Karen): I don't think it actually has any impact at all on the validity ratings. It's really …

(Ron): And it's (not) -- it's -- I mean, other than a (exercise), which I think we should do because I think (Sherri) has raised an important question. Other than having an impact on fine-tuning their response to the measure, it's not going to change what ultimately happens because we've got four high ratings on validity.

(Karen): Well, we have to know what your ratings are for reliability. So if you guys voted right this second and you said, "Hey, (Sherri's) convinced me what they've done isn't quite enough or quite right." and a bunch of you vote insufficient on reliability, then the measure goes down. So it has to go through …

(Purna Renee): And (Ron), just some clarification. I think you're mixing up what's happening in this measure, what happened in the last two measures. Those …

(Ron): Yes?

(Purna Renee): … two are not instrument-based, so we don't require (both) …

(Ron): Oh, (OK).

(Purna Renee): … (measure scores and these elements).

(Karen): Yes, there you go.

(Purna Renee): So with that one, we do allow that concession to do validities. But (with) …

(Ron): Got you, yes.

(Purna Renee): Yes.

(Ron): I see it now.

(Purna Renee): (So just to make sure you understand the difference).

(Karen): Thank you, (inaudible) …

(Ron): Never mind.

(Karen): I didn't understand why (Ron) was OK as long it went through on validity. So apologies, I wasn't following you, (Ron).

(Ron): I am temporarily back to insufficient on reliability. I'm sufficient in doubt, so my rating is rather not high but insufficient.

(Sherri): This is (Sherri). (Karen), I'm sorry, I have to get off for now. So …

(Karen): OK.

(Sherri): … I will look forward to your -- what your decisions are. Take care.

(Karen): OK. Thank you, (Sherri). So I think what we're going to do, unless the team really gives me a dirty eye, I'd like to talk about this on our next scheduled call. In the meantime, we can see if the developers have any extra things that they could provide.

And I think we also -- I may tap (Sherri) to see if she can write up something, or provide some formulas or something like that to make sure that everybody's completely on the same page.

And one of the reasons I want to do this is this methodology and the question of methodology has come up multiple times and it's going to continue to come up.

(Ron): Yes.

(Karen): And we've talked about it on our monthly calls, but I don't think -- I think, to be honest with you, I think we need to have that discussion in context of a measure. And without being able to do that, it just makes -- it's harder to do, theoretically, I think. So we'll use this as a learning opportunity for all of us.

(Jack): Yes, right.

(Ron): And I apologize for forgetting that that was an instrument-based measure for a couple minutes.

(Karen): Oh, no worries. I'm just glad (Purna) figured out the -- what was going on. All right, so what we will do is we will get back to you before the next call. If there is a possibility of doing something over e-mail, we'll let you know too.

But I think -- I think we probably need to have this call, get it all documented, that sort of thing.

But we did really well. I think this was a successful call. I hope you guys thought so as well. If you have ideas for things that we could do better, let us know. Hopefully your voting is working, if not, let us know. Anything else we need to?

(Purna Renee): No, just seeing -- we received all the votes for 0753 and if you want to turn in your votes for 2456, that would be great. Again, you're only voting on -- (I'm sorry), validity for 2456. We're good to go for 1716, 1717 and we're going to hold for this -- the possible vote on 3450 until we get you a little bit more information.

(Jack): So I voted twice, you've got all my votes?

(Purna Renee): (All right).

(Karen): All right.

(Ron): Yes. Have mine on 2456, too.

(Karen): (Thank you) …

(Ron): (No, you should have).

(Karen): Thank you for your patience as we try to work out a process that works better than what we did last year.

(Jack): This is better.

(Karen): Do you think it's better?

(Jack): Great fun. I'm learning from my colleagues on this. So I want to thank everybody else on the panel for the work that they did thinking about these measures.

(Karen): These were big lists, yes. So thank you, guys. And we'll be getting back to you very soon. Have a great rest of your afternoon.

Female: Thank you, bye-bye.

(Purna Renee): Thanks.

Operator: This concludes today's conference call. You may now disconnect.

END