NATIONAL QUALITY FORUM

Moderator: Sheila Crawford
March 15, 2019
1:52 pm CT

(May): Hello, everyone. This is NQF staff and we'll be starting shortly. I think a lot of people are just dialing in now, so we'll start in one to two minutes.

Man: (Unintelligible)

(May): Good afternoon, everyone. This is (May) from the National Quality Forum and we'll be starting our Subgroup 3 Scientific Methods Panel Conference Call today. And I'll just start out as people are still dialing in, I'll start out with just quick housekeeping remarks and then I'll turn it over to Ashlie Wilbon who is a Senior Director to perform roll call and (DOI).

So, to let everyone know, the discussion guide was actually sent to Subgroup members yesterday. This document will guide us in our measure discussions today and we will follow it with the order presented on the document.

Consensus was not reached for four measures presented on the discussion guide and are subsequently slated for discussion today. All other measures will not be discussed during today's call unless the member of the subgroup

would like to poll that measure. If you choose not to discuss additional measures, the decisions from your preliminary analysis will be made final.

In that same email, with the discussion guide, there was a link to a SurveyMonkey. So, we ask the subgroup members to pull that survey up now and cast your votes on reliability and/or validity at the conclusion of each measure discussion.

Staff will prompt you to cast your votes when the timing is appropriate. Timing today is limited to roughly 30 minutes to discuss each measure, although we would like to come to consensus on all four measures today. We do have a follow-up meeting scheduled on Tuesday, March 19th at 10 a.m. Eastern Time to discuss any outstanding item.

Finally, I would like to note that this is a public call. Developer representatives may be on the line to answer questions from staff or from panel members. However, there will be no opportunity for public comment.

For recordkeeping purposes, we ask that you please say your name each time you provide remarks.

And now I'll turn it over to Ashlie to perform roll call and DOI. Ashlie?

Ashlie Wilbon: Thanks, (May).

Welcome everyone and thank you for joining us. So, I'm going to combine the disclosure of interest with our roll call for today for the method panel members. So, just bear with me as we make it through all of the preamble that goes with that.

So, you received a disclosure of interest form from us before when you were named to the committee and then we send one every year, kind of an overall disclosure of interest. And then we also send more specific - measure-specific disclosure of interest form that asked you about the specific measures you're going to be reviewing, whether or not you have any relationship to the measures that are specifically under review or any related or competing measures that we may have identified.

So, between these two forms, we do ask you quite a lot of questions in terms of your degree of involvement with the measures under review. So, in the interest of transparency, we do ask on calls that we review measures to have those review in measures to early disclose any interest or conflict of interest they may have.

So, today, we're going to ask you, like I said, to early disclose any information that you believe is relevant to this committee, specifically in the measures you reviewed as a member of this subgroup and any related or competing measures. So, we don't need necessarily some of your resume, only any particular activities, research grants, and full paying - paid or unpaid activities latest in measure development for the measures that are specifically under review.

So, this group reviewed 10 measures (unintelligible) to readmission measures. One for Facility Seven-day Risk Standardized Hospital Visit Rate for Outpatient Colonoscopy and another for Hospital-Wide 30-day, All-Cause, Unplanned Readmission. And then, there were eight cause measures around various conditions for PCI, lower extremity chronic critical limb ischemia, pneumonia, cataract removal, colonoscopy, knee arthroplasty, pneumonia, (unintelligible).

So, several measures that you reviewed and a couple reminders as well before we get into the roll call. You sit on this group as an individual. You don't represent the interest of your employer, anyone that may have nominated you.

And just because you disclose, it does not mean you have a conflict. Again, we just do this in the spirit of openness and transparency so we'll go ahead and get started. I'll call your name, please let us know if you're here and if you have any disclosures of interest.

Karen Joynt Maddox?

Karen Joynt Maddox: Present. In the middle of a sneeze.

Ashlie Wilbon: Thank you.

Karen Joynt Maddox: I do not have any direct disclosures. I do contract work for the - for HHS with the Office of the Assistant Secretary for Planning and Evaluation so I interact with CMS and Acumen and folks like that through that work but nothing related to these measures.

Ashlie Wilbon: Thank you. Jenifer Perloff?

Jenifer Perloff: Sorry. Same problem. Hi.

I'm here and I do want to make sure I'm clear. My research work is involved developing and episode group or for Medicare. Acumen was our evaluation contractor during that work but I was not involved in any way, shape, or form in the development of these specific measures but I'm sort of intimately familiar with this space.

Ashlie Wilbon: Thank you, Jen.

Ron Walters? Are you there, Ron? Are you on mute? Okay.

Christie Teigland?

Christie Teigland: Yes. I'm here and I do not have anything to disclose.

Ashlie Wilbon: Okay. Thank you.

Jack Needleman?

Jack Needleman: Yes, I'm here and no disclosures.

Ashlie Wilbon: Okay. Thanks, Jack.

Susan White?

Susan White: Yes. I'm here. No disclosures.

Ashlie Wilbon: Hi. Thanks, Susan.

I'm just going to go back to Ron. Ron, are you there?

Ron Walters: I'm here now. Sorry. I'm in clinic …

(Crosstalk)

Ashlie Wilbon: Welcome.

Ron Walters: Had to run out of the room.

Ashlie Wilbon: Okay. Welcome. Thanks for joining us.

Ron Walters: No problem.

Ashlie Wilbon: Do you have any disclosures to share?

Ron Walters: I have no disclosures to share.

Ashlie Wilbon: Okay. Thank you, Ron.

Ron Walters: Okay.

Ashlie Wilbon: So, again, thank you everyone for disclosing and letting us know that you're here. If at any point, during the discussion you feel that anyone is speaking or acting on a bias manner, please let us know. You can email us or staff and we will do our best to address any concerns that may arise.

Are there any questions based on the disclosures that (were named) today? Okay. Great.

So, today, as (May) mentioned, we have four measures that are slated for review. Measure 25, 39 which is one of the hospital-based visit rate for outpatient colonoscopy and then three of the cost measures, one for elective PCI, revascularization and pneumonia. There were four measures that passed based on the preliminary evaluations that were submitted by method panel members.

These measures are not slated for review on this call unless someone from the methods panel decides - would like to suggest that they are pulled for discussion. Those measures are 3495 which is the Hospital-Wide 30-day, All-Cause, Unplanned Readmission; 3509 which is the Cataract Removal with Intraocular Lens Implantation was a cost measure; the Screening and Surveillance Colonoscopy Cost Measure and the Knee Arthroplasty cost measure.

There were also two measures that did not pass the initial preliminary analysis by method panel members. That was 3514, the Intracranial Hemorrhage or Cerebral Infarction measure and 3515 STI, I'm sorry, ST-Elevation or STEMI with PCI.

So, again, I just want to make sure to - that everyone was oriented with what we're going to be discussing and also to give methods panel subgroup an opportunity to call out any measures in addition to what we have already slated for discussion. Is there - are there any requests for other measures to be pulled for discussion?

Ron Walters: This is Ron. Can I just ask a question?

Ashlie Wilbon: Sure.

Ron Walters: This can be of a general question, okay? Not specifically a measure-oriented. With so many measures that fell basically under the same category of cost and efficiency, many of which were done by the very similar, if not the same methodology, okay? So, it was almost a completed methodology for certain some of them.

Could I hear some feedback about what people notice - what their major contentions were about the ones that did not pass very the ones that did within the cost of efficiency group? Was it the …

(Crosstalk)

Ron Walters: … quality of the analysis? Was it the type of analysis, et cetera?

Ashlie Wilbon: Right. Right. Ron, can I hold that question for just a second because we're …

Ron Walters: Sure.

Ashlie Wilbon: … I was actually going to get to that in just a second and we can talk about kind of the best way to go about evaluating those measures because, you're right, there are a lot of kind of the underpinnings and the construction of those measures and the testing approaches were very similar.

Ron Walters: Okay.

Ashlie Wilbon: So, there are nuances with each of those conditions where some of the testing results, even though they applied the same method, the results were different. So …

Ron Walters: Yes.

Ashlie Wilbon: But we do have a couple options for how we can go about evaluating them …

Ron Walters: Okay.

Ashlie Wilbon: … and talking and talking about them. We can do kind of an overarching discussion and then dive in to each measure individually.

Ron Walters: Thank you.

Ashlie Wilbon: So, yes. So, we'll come back to that in just a second.

Ron Walters: Okay.

Ashlie Wilbon: And I think we'll have a discussion as a group about the best way to go about that.

Ron Walters: Okay. Thank you.

Ashlie Wilbon: Thanks for bringing that up.

I did also just want to point out that this call is the first since we've started the subgroup calls with method panel members that we're inviting developers to answer questions and be available to method panel members to respond to, any concerns that come up during the discussion.

So, I did want to just take this opportunity to check in to make sure that developers for these measures that will be on the (review), they are on the phone from the developers from Yale and Acumen.

(Crosstalk)

Ashlie Wilbon: Okay.

Woman: Our team is on the phone.

Ashlie Wilbon: Hi, (Liz).

(Elizabeth): Hi.

Ashlie Wilbon: What about Acumen? Is there someone …

Sri Nagavaruko: Yes. This is Sri Nagavaruko from Acumen. We have a team on the phone.

Ashlie Wilbon: Okay. I'm sorry. Can you say your first name again, please?

Sri Nagavaruko: Sure. Sri. Just S-R-I.

Ashlie Wilbon: Okay. Thank you.

Okay. Thank you, guys, very much. So, I'd also just wanted to point out a quick process point for the cost measures. We are going to discuss the colonoscopy outpatient measure first. But just as a quick point of process as we make it through, the valuation process for the cost measure is slightly different than some of other committees mainly because, actually, very much likely readmissions committee, the cost committee was not - is not a clinically-based committee. The experts that are comprised on that standing committee are - there are some positions who, obviously, are clinical experts but there are also a lot of economist, methodologists and other background and expertise on that committee.

And so, oftentimes, when we do get clinically focused episode-base measures, cost measures for review, we do employ that help up clinically - clinical expert test that are convened to provide the standing committee with clinical guidance as well on the clinical aspects of the measures.

So, for those measures, in addition to the methodological input that we get from the message panel, we will also be convening various clinical tests to review the clinical aspects of those measures.

So, those tests have already been convened and are place. And so, regardless of the results of the method panel results of the voting, all of the measures will still go forward to the technical expert panels. So, all (managers) of those measures will still get the clinical input.

And based on whether or not they passed the method, the methods panel results or not, we will then determine which measures go forward to the committee.

So, I just kind of wanted to point that out at some kind of a nuance to our process with these particular measures. And it will be -- as we begin to receive more of these episode-based measures -- it will be a process that we'll probably continue to use in order to make sure that we get the adequate expertise and input on these measures as they make it through the process. So, I just wanted to make sure that folks were aware of that and as measures kind of make it through the process.

Jenifer Perloff: Can I just …

Ashlie Wilbon: Are there any …

Jenifer Perloff: … quick question on that?

Ashlie Wilbon: Sure.

Jenifer Perloff: So, this is Jen. So, what - the implication of that is that the clinical specifics of the measure should not be as much of our concern here today. That there will be someone paying attention to some of those clinical details.

Ashlie Wilbon: Yes. And certainly, to the extent that you have that expertise or you have concerns, we certainly will document those and include those as part of the feedback but there are clinical experts as well. So, if you're maybe not as comfortable about some of the - or have - maybe that your expertise is not in the clinical space, you can (rest) but there are - there will be clinical experts that will be looking at of as well.

Jenifer Perloff: Excellent. Okay. Thanks.

Ashlie Wilbon: Yes. So, with that, why don't we go ahead and jump in to 2539 and then we'll have some discussion about the cross measures when we start that group of measures.

Okay. So, for 2539, again, it was one of the measures that where consensus was not reached. It's past validity with a moderate score, reliability was the criterion where the consensus was not reached. I just want to bring a few things to your attention as we dive in.

So, this measure was reviewed by the methods panel in the fall cycle. The fall cycle was the first submission of the measure. It was previously endorsed so this was - this is the first maintenance review, if you will. So, it is a maintenance measure.

And I won't rehash all of the details in terms of the specs about the measure. You have that in front of you. But I do just want to highlight a few things

before we dive in to the discussion and kind of highlight some of the things that, I think, should be a focus of discussion by the methods panel.

So, I'm (going to kind of) skip down (unintelligible) or reliability and validity. So, with reliability, we ended up with one high score, two moderate, two low, and one insufficient which (unintelligible) in that consensus not reached (soon).

For the reliability testing, they submitted signal to noise testing for the measure score. They presented data test for - based on two samples of data. One for three years of data, another for one year of data and the results for those are (adhered) on the discussion guide. For the three years of data, the median score was 0.814 for the hospital outpatient department. Out - yes, outpatient departments.

And then 0.893 for the (ASCs). And then for one year of data, the median reliability was 0.593 and then 0.735 for ASC. So, for this particular criterion, there was concern over the reliability testing results particularly for the one year - for the results presented for one year of data. And also, a note that the measure is reported using a single year of data.

And so, obviously, I think the discussion is going to be there. I just wanted to point that for the last submission to the methods panel, the measure past reliability but not validity. So, we have a bit of split of the results for this particular evaluation.

So, I just want to make sure that folks kind of understand potentially the differences between the submission that they submitted last cycle and for this cycle. And so what I - from what we can tell so far, you may just want to

check in with the developers the difference and what they submitted last cycle and this cycle is last cycle, they also submitted split (processing) results.

So, that was not included in the submission. And just to point out, I know that one of the reviewers had kind of question why that testing was not submitted and in NQF, we don't have any requirement on submitting both signal to noise and split half or either/or, signal to noise is an acceptable method of reliability testing that we do (assess). And so, we just wanted to make sure that the voting wasn't (requested) of them not having submitted (the split half) testing because that's not necessarily a requirement.

The evaluation should be based on what they have submitted which is the signal to noise analysis. So, with that, why don't I just pause there and see if there's any questions and maybe open it up for discussion for the panel for that (element).

Jack Needleman: This is Jack Needleman and just, I think, one of the themes that's going to be all through today's discussions is whether the characterization of reliability of that has been presented and that often (aside) and then the literature are, in fact, the standards we want to apply for levels of the liability for purpose of use.

And I will just say that in general, I find the standards from the literature when we look at the actual implications for classification, those standards tend to be lower than the ones I would apply in terms of the standards for liability and that's going to run through all of my discussion today.

But I think for the committee, in general, thinking through that issue and getting some committee-based standards for where we think the standards of reliability or correlation or stability of rankings is (unintelligible) that the

committee is active and front as the committee, (separate part) from the review of the individual measures.

Susan White: This is Susan White. I had an issue - my reliability issue was actually beyond that and was in the way that they sort of projected the three-year - they use one-year data to calculate the reliability and then projected it to what it would be with three years of reliability. And I think that assumes a - there's a big assumption around the stability of the measure throughout the years that I - that was what sort of made me pause on whether the reliability was acceptable or not and not even looking at the level of the score, but just the methodology and the way they applied it.

Ron Walters: And this is Ron. Susan, that's exactly what I wrote in my assessment also. And that probably, the only reason that I called it low reliability rather than moderate was exactly that assumption and the way the calculation was done three years versus one year.

Jenifer Perloff: This is Jen. I think I confess that I somewhat ignored the three years because it's a one-year measure and it seems like the one-year results were the ones that mattered. So, I was more in Jack's camp around sort of the absolute value of the reliability statistic.

Jack Needleman: Yes. And this is Jack again. If the measure's going to be based upon one year of data, then the three-year reliability - the question I asked in mind was is this going to be based upon three years of pool data? Is it going to be based upon one if it's one and then the three-year figure is irrelevant?

But I appreciate if it is. If it's three actual years, that's not specified here. So, if it's actually a one-year …

(Craig Parzinsky): Hi. This is (Craig Parzinsky) at Yale (corner). I was wondering if we could clarify?

Ashlie Wilbon: Yes, please. Go ahead.

(Craig Parzinsky): Sure. So, first of all. To address the split half reliability calculation, we did include that in our original submission and that's what you're referencing. We did decide to poll that and I think that was exactly for the reason that you're suggesting that we have to project that data because we, at the time, did not have what would require six whole years of data (unintelligible) three years what sample of reliability. We were able to acquire three years of data to do the signal to noise and that's what was submitted with this submission.

In terms of what is going to be used for (unintelligible) reporting, CMS has recently released in their regulation that they will be using a three-year data period which is also partly the reason why we submitted the additional three years of signal to noise.

Ashlie Wilbon: So, this is Ashlie from NQF. I'm sorry. I didn't catch the name of the developer from Yale that was speaking. But …

(Crosstalk)

(Craig Parzinsky): Sorry. It's (Craig Parzinsky).

Ashlie Wilbon: Right. Thank you.

Can you -I just want to clarify. So, the specifications that you submitted for the measure, do they also request that the measures reported and aggregated over three years?

(Craig Parzinsky): The specifications, I believe, that were submitted likely reflect mostly one year data and three years of data may - we may not been able to run everything at that stage when we are trying to meet submission deadline.

Ashlie Wilbon: Okay.

(Elizabeth): Yes. And, Ashlie, this is (Elizabeth). I think we're talking about the data and (specifications). We could formally specify this as three years. CMS was pushing that decision to rule making and finalizing while we were (giving you stuff). So, the timing is just - it's not completely aligned, which I'm sorry, it makes it confusing.

But it is finalized in rulemaking for three-year sort of data for reporting. And so, the measure could be - I'm (moving through) it. I'm making a - I'm raising a question because it's really CMS' decision. I'm not sure if they're on the line.

But the measure could be formally evaluated for re-endorsement as a - with three years of data. That's part of what you want to see specified as the quote-unquote specification.

Ashlie Wilbon: Yes. I think the concern is that we want to make sure that the testing - if you've done testing to demonstrate that the measures are reliable as certain kind of sample size or data, we need that to also - the testing should reflect the specifications. And so, those need to align.

What we don't kind of want is the (decisions) made for recommending based on one year of data and then the measure gets used another way or vice versa.

And so, to the extent that any decisions that we make just needs to be consistent that voting on reliability, for example, we're voting based on the results for three years and that the specifications will be aligned to support that as well. So …

(Elizabeth): And can I ask you a question? This is - sorry. This is just to go to what you're voting on or how you - so, in a different data, that was a different - I mean, if you're specifying - if you're evaluating (for use) in this population for these facilities, the hospital outpatient departments and ASCs, if you're - if the endorsement is specific for those settings, it absolutely will be in three years of data.

If you're saying could this measure be used in a different setting, then you might need five years of data or one year of data, it would depend on the outcome rate and the variation and all those things because you would have different reliability when you run the test.

So, I guess I'm - could you share - do you need - I mean, if you're - if this committee is looking at it (through) endorsements for this use in each setting, the assumption should be because it's finalized in regulation three years of data.

(Crosstalk)

Susan White: The testing wasn't on three years of data, though, right? It was on one?

(Craig Parzinsky): The reliability data in three years of data and the reason that that was available is because we performed that analysis for the regulation writing but - and we were unable to kind of run it and all of the results in the three years of data. I would argue that one of your data does have (proficient) reliability do,

however, much of the literature that's out there suggest that a value of 0.59 or higher is in the moderate to high range of the vast - I don't think it should be a major sticking point but I did want to be able to clarify why the three years was in there and try to hold off any prolonged conversation questioning whether or not it was in regulation or not.

Ashlie Wilbon: So, this is Ashlie again. I just want to make sure I understand (Liz's) question and then I'll turn it back over to the method panel.

So, the description of the measures specifically also points out the hospital outpatient department and the surgical center. But what you're saying is that the measure could be used for other settings? I guess I just want to clarify for my (own understanding).

(Elizabeth): Yes. I mean, it's embarrassing and I apologize that I don't know that - (unintelligible) how much I know about NQF and how involved I've been. But I'm just saying whenever you endorse a measure, right, it gets endorsed with a particular dataset. So, you're endorsing the measure in that context, right?

I mean, once it's endorsed, people use it - people may use it in another way then they may adapt it and apply (it back to). I just don't know how tight that endorsement is. Are we meant for this setting only, right? But I would just say that in - I want to just - and I don't want to (hang up) the committee on this at all. I just - for this setting, for this use, it's three years of data. That's finalized in rulemaking very recently.

Ashlie Wilbon: Okay. And I would say yes that the endorsement does - should match the setting, I mean, the specifications and the measures we endorse. It should match the kind of the level that it's endorsed based on how the measure

specifies. So, for the setting, the level of analysis, the measurement period, all of those things are (baked in) with the measure.

Now, what you're referring to and how people kind of take the measure and apply it in different settings, that's something that kind of out of our hands. So, a lot of - to a large extent. So, I would just say the committee - the panel and the committee is going to be looking at the measure as specified and kind of what happens after it's endorsed potentially that's you don't have a lot of control, all we can say is this is how the measure - we endorse the measure for use in the specific conditions based on the specification.

So, with that, I'll just kind of hand it back over to the methods panel to see if there's any other clarifying questions about the three years versus the one year.

Jenifer Perloff: This is Jen. I would just say I think it's a point that I appreciate that you may expand or contract the time that you collect the measure across and it - so, in different situations, you may need to add more years of data. How that gets reflected in the form and whether developers should show us one, two, three, five, multiple years, so we can understand the boundaries of that. I just want to know if that's an important point and I appreciate that.

So, here we have one and three and that gives us the contours of that. (But right) that the measure has to be sort of endorsed for a specific time or range of time, maybe multiple times.

(Crosstalk)

(Elizabeth): Well, (here) I just want to clarify. It's (Elizabeth) again. I just - what CMS is - so there was - this is just a tiny thing. But what we would really like NQF to

focus its review on is three-year timeframe. That's what the - that's the re-endorsement that it's most relevant to the current use of a measure.

Ashlie Wilbon: So, this is Ashlie. That said, are there - I know that - I think it was Susan that had initially pointed out some concerns about the way that the analysis for the three years of data is done is that have those concerns been addressed or is that …

Susan White: So, this is Susan. I think some of my colleagues convinced me that it might not have been relevant since we had one-year but now I think it is relevant since (we're supposed to be evaluating) on three years although I might be confused.

Ashlie Wilbon: No. I think that's an accurate characterization, Susan. I think based on what the developers have shared now about how the measure will actually be used, I do think that we need to - we may need to do some reconciliation with the specifications on the backend. But it sounds like the measure will be used for based on three years of data. So, that should be, I think, the focus of the discussion at this point, it sounds like.

Jenifer Perloff: Can I just clarify the three years is not a simulated data, it's all three real years?

(Craig Parzinsky): Hi, this is (Craig) (unintelligible). And yes, this is three years of whole data. The split sample is what needs to be projected because you would require (unintelligible) to do a split (year) for three years (each). And so, that why we're going to (reason at that) because of (unintelligible) difficult to get six years of continuous data.

Jenifer Perloff: Yes. Got it. Okay. Thanks.

Ron Walters: So, just to be fully clear. The reliability figures that are presented here are from testing on three years of data?

(Craig Parzinsky): Right.

(Crosstalk)

(Craig Parzinsky): Yes. The point …

(Crosstalk)

Ron Walters: They're not a simulation or a projection or modeling based upon using one years of data to try to simulate three years of data.

(Craig Parzinsky): Exactly right. The 0.81 and the 0.84 values are from real data.

Susan White: So, I'm looking at the - I'm actually looking - I'm looking at the measured - I think I'm looking for measure testing (form). Yes.

Sorry. Never mind. Go ahead.

The actual measure testing form and it says in Section - I got to find it again. Sorry.

Section 2A2.3, the (unintelligible) reliability score, parenthesis, (projective for three years data), end parenthesis. And then below, it says using a single year of data 2017, the median reliability is 0.5 times three. So, how am I - I'm getting confused and I need help.

Ron Walters: No, Susan. That's exactly the same phrase that I put in my comment, too, under Question 7 and I still haven't heard that explanation either as to why that is a direct pull from the submission form that implies that these three years of data was a derived result. And see, I think that's what's causing a lot of the confusion.

(Craig Parzinsky): This is (Craig) at (Core) again. I'm wondering if that's a type of we will do some quick back and work here just to confirm that. This may be from left over from the prior submission. So, (we will check that).

Ron Walters: And this is Ron. We have seen that happened before. So, it's possible. But Susan's right. It's exactly in Section 2A23 of the submission form.

Susan White: It's Page 7 of the testing form.

(Elizabeth): Great. We'll look at that. I'm sorry about that confusion. If we - this is a second submission (unintelligible) and we'll fix it. So, I don't know if you want to move on and - (Craig), is that something you can look at in real time?

(Craig Parzinsky): Yes. We're going to try (digging) to that right now and get back to (unintelligible).

Susan White: Okay.

Ashlie Wilbon: So, this is Ashlie. I just want to clarify - so, I know verbally, you're telling us that the data has not been projected. Are you pretty confident in that? Should we - is it that you think that the submission form needs to be edited or is there a possibility that it is actually projected. I just want to make sure, like, what (Craig) is checking on, is it checking on that it's a typo or checking on

whether or not the data actually was projected or not. I just want to make sure.

(Craig Parzinsky): I am checking on the typo and we did run this for the rulemaking process. So, we're just confirming that it is a type carryover.

Ashlie Wilbon: Okay. Okay.

Why don't we do this? So, let me just jump to validity really quickly. And then before we have you guys submit your votes, we'll just check in with (Craig) again on the reliability and see whether or not there's any further discussion that needs to happen if that's okay with folks.

Okay. So, for validity, so for this round of the review, we had zero high, four moderate, one low and one insufficient. So, that essentially put us out of moderate rating for this sub criterion.

So, a couple things that I wanted to point out. This was a maintenance - this is a maintenance measure. And so, technically, our rules - our requirements for maintenance measures is that for - by the time the measure comes back for maintenance review, that there has to be some empirical analysis to demonstrate validity.

So, initially, they had done space validity which was deemed to be accessible and because there was a challenge in demonstrating empirical validity, and in effort to find other measures to correlate with for kind of a (concept) validity assessment, the developer submitted a rationale in their submission for why they were having difficulty being able to demonstrate that empirically. That - the demonstration of empirical validity is a requirement but we do allow the

submission of a rationale if their unable to meet that requirement. And so, the question then becomes is that rationale, satisfactory for the reviewers?

So, based on how the ratings were submitted for this criterion, it is reflected at - in that - as you reviewed this, that that rationale was acceptable. Generally, when a criteria has been voted on by the methods panel, we do not kind of rehash it. But I just wanted to - we wanted to make sure that looks clear to folks and their voting to make - to see if there wasn't any clarity whether or not we need to have a discussion about that and determine whether or not there was any questions about that particular rationale.

I will also note that in the previous evaluation of this measure and last cycle for the methods panel, this actually was the tipping point as opposed to the reliability. So, again, I mentioned things how kind of - things have kind of flipped between the two cycle. So, again, I just wanted to bring your attention to that and see whether or not there was any - you guys felt that there was (anything we need) to discuss or are you settled and you're okay with the rationale?

Susan White: So, this is Susan. I brought up the point that there are some other similar measures that they can use for external validity. So, I did have an issue with their statement that it couldn't be done. I could use any of the seven-day hospital visits after surgery performed (at ACS). It should be correlated. It should be directionally correct or directionally aligned. So, I was kind of surprised that that was the rationale for not having some sort of - something beyond case validity.

(Crosstalk)

(Elizabeth): Can I make one quick comment? It's just - the ambulatory surgery center setting, in particular, there's - these are unique - this tend to be especially specific facilities. So, we actually just didn't feel comfortable saying that these could even - they're just completely - as soon as they're - they're usually physician owned. So, (run) by a group gastroenterologists. So, there's two gastroenterologists then three ophthalmologists.

There's so little overlap that we didn't feel like going to a whole another procedure group because the (unintelligible) overlap it all, the (group) of procedure (soon) overlaps it all. (Unintelligible) impression that we wouldn't know what to do with the result, that they were correlated or uncorrelated. We didn't really feel like a (unintelligible) they should be correlated.

Jack Needleman: This is Jack. I'm reasonably comfortable not having a correlation with another measure here. It seems to me what we're measuring is trying to get some implication or direct measure of unplanned complication problem that requires a hospitalization. Among folks who may not have a lot of hospitalizations for other things.

So, I'm willing to buy that. I'm more concerned about, actually, given the nature of the - that this is (intended) to capture unplanned readmissions because of complications, some lapse in the care whether the specifications, in fact, do that, and as I read the report, there was actually this statement of concern amongst some of the TEP members that if you (saw) a result on the colonoscopy that would want you - you'd want to put the patient in the hospital to follow up on that, that's not unplanned in the sense of it's a complication, something was (then) anticipated from - in the treatment, something didn't go quite right.

And I didn't see any discussion of how that concern was dealt with in the specifications either by saying it seems to be de minimis so we can ignore it or we've dealt with it in the specifications by exclusion cases that look like X or not counting cases that look X.

So, from my perspective, I'm more concerned about not the lack of correlation with other measures but with whether they got the specifications right.

(Elizabeth): Ashlie, you just have to give me guidance on (this time) whether you want me to - on these comments. I don't want to over …

Ashlie Wilbon: Yes. I wanted to just give a - if there are a couple other - is there anyone else on the - for the methods panel, any other thoughts on this and then a list of - we'll have you give a rebuttal. Any other comments?

Is there - let me just ask this. Is there a need to rehash this or do you feel like your votes reflected your understanding of them submitting the rationale for not having empirical validity testing? That's really the question.

If you feel settled on that, there's no need to rehash it. I guess that was really the question to make that clarification and make sure that that was the common understanding and then that is - if that was the common understanding, then there's no need to kind of - to dive into that discussion.

So, okay. So, while you guys are thinking about that, I don't know if silence is agreement or not, but, (Elizabeth), why don't you go ahead and add any additional commentary you think might be helpful?

(Elizabeth): Sure. Thanks, so much.

So, the issue of could a hospital visit be not unrelated to a quality event, be related to ongoing care, for example, for the patient, that was really central to the full deliberations on our expert panel and public comment. And we do have specific exclusions in the measure that address that, for example, we don't include patients with (unintelligible) disease or diverticulitis because it was unclear if they visited a hospital within seven days and we couldn't - we didn't have an algorithm in claims data we could use to start with.

That's because they had a - the reason they had the colonoscopy in the first place was diagnostics (unintelligible) and then their conditions just got worse. We did some data analysis around that and we weren't confident that when we saw (an admission) from that group of patients, that we were seeing an unplanned visit even after we (pulled) our plan, what we usually go out for plan procedures.

So, that was intensely focused on by our expert panel and we refine the measure along the lines deliberately. And in the end, there was - strong face validity support from the measure. So, if we discussed that in the application, I'm not sure exactly which page it's on but that helps to provide a little bit of context on the clinical approach to ensuring the measure that has that face validity.

Ron Walters: And this Ron.

(Crosstalk)

Ron Walters: I just want to clarify something you said earlier. Face validity in this particular circumstance is good enough, right?

(Crosstalk)

Ron Walters: Were you referring to me? Ron, who was that directed to …

Man: Who is that directed to, Ron?

(Elizabeth): Ashlie, can you restate? Because my understanding of what you said before was that we need to try do external validation and (period) validation. But it's not required for re-endorsement and that face validity is still relevant here (unintelligible).

Ashlie Wilbon: It is still relevant. It is - I mean, it is required essentially, unless there's a rationale to demonstrate why it cannot be done. So, that's really the question of whether or not the rationale is acceptable. Yes.

Ron Walters: Got it. Yes. Okay.

Jenifer Perloff: So, the question is (if we're comfortable) is on the table, I guess, the low - it is important for the folks who felt that this is low and insufficient on validity whether they have remaining concerns, right? Isn't that sort of …

Ashlie Wilbon: Not necessarily.

Jenifer Perloff: Okay.

Ashlie Wilbon: I mean, even with those folks staying where they are, the measure, still, would be at a moderate. And so, I think it was more for the folks who voted moderate, was that your understanding of what you were evaluating, the rational for not submitting empirical evaluation versus the - like them having face validity?

Ron Walters: This is Ron. I was moderate and yes.

Ashlie Wilbon: Okay.

Jenifer Perloff: Jen as well.

Ashlie Wilbon: Okay. Okay. So, it would be helpful to hear from others who voted moderate just to make sure that we're kind of checking our - crossing our Ts and dotting our Is. But it sounds like from the folks - from the discussion so far that folks were okay with the rationale, no need to rehash that issue and then we can circle back to reliability and then, hopefully, vote and close on this discussion.

Is that acceptable to folks? If you could just say, like, yay or nay, to let me know. We're settled with the validity.

Karen Joynt Maddox: Yes. This is Karen.

Ashlie Wilbon: Okay.

Ron Walters: I am. This is Ron.

Christie Teigland: Yes. This is Christie. I'm good. I was moderate. I'm fine.

Ashlie Wilbon: Okay. Thank you. That's very helpful. Okay.

So, (Craig), I don't know if you're still there but if you could give us an update on reliability and to just refresh from the discussion we had about five minutes ago, we were wanting to verify that that the data that was - that the three-year data sample that was tested was, in fact, three years of actual data not projected or simulated data and that the evaluation for reliability when the

committee or when the panel vote shortly should be based on that three years data and there are kind of feelings about whether or not that approach and the way the testing was done and the results for that particular data sample was appropriate. Correct?

(Craig Parzinsky): Yes. Yes. Thank you, guys. I didn't have the chance to look at the result and it doesn't look like it was a typo that we carried over the same section from last time. The results are publicly available right now in the rule. And so, I'm going to read them out loud for you just for the three years of data.

For ASCs that the reliability was 0.87, so pretty close to the projected data and then for HOPD, that was 0.75 which is, again, fairly close to the projected data. And so, that's in the real data and available in the final rule that we can send to you guys up for this meeting. So, again, apologize for the failure to update the testing plan properly there.

I, myself was confused when it said projected because I knew that this was already released in the final (unintelligible). Apologies for that.

Ashlie Wilbon: Craig, this is Ashlie. Can you repeat that one more time? I just want - I didn't get a chance to write it down, sorry.

(Craig Parzinsky): Sure. Do you want the reliability results?

Ashlie Wilbon: Yes, please.

(Craig Parzinsky): Yes.

(Crosstalk)

(Craig Parzinsky): For three years of data, in the final rule for ASCs, is 0.87 and for HOPDs, it is 0.75.

Ashlie Wilbon: Okay. So, those are numbers that the panel should really be based in their valuation on.

(Craig Parzinsky): Yes.

Ashlie Wilbon: Okay.

So, a couple things here. I would like to go ahead and close the discussion out so we can move on to the next set of measures. A couple of just clarifying points here.

One, regardless of how the panel votes, we're going to need some update - we'll have to work with Yale to get some updates to your submission forum before it moves forward just so we're not kind of perpetuating confusion. And to the extent that we can make sure that the specifications match the measurement period kind of matches the testing that was performed and that three-year issue is resolved.

For our issue, for our voting today, I want to make sure that the panel is clear that the direction now is to have you guys vote on your evaluation of the reliability testing for based on three years of data at the ASC level which was 0.87 and for the HOPDs which was 0.75. Is there any other discussion that you'd like to have on that before you vote?

Okay. It doesn't sound like it. So, what I will do - what I will ask then is, (May), sent an email out yesterday with the link to the SurveyMonkey. So, we will ask that you kind of locate that survey and open it up and submit your

votes only for reliability for this measure, please. And we will make sure that gets submitted.

So, we're going to go ahead and move on to the next set of measures, the cost measures. And kind of shift gears here a little bit. I did want to just make a few comments. Actually, very similar to the discussion or the points that Ron raised earlier on.

These measures are structured or constructed very similarly but for a - this is for different clinical areas. There's a lot of - there's going to be a lot of (unintelligible) that will be applicable to many or all of these measures around reliability, around risk adjustment, and some of the exclusions and so forth and even - it was validity.

So, kind of keeping those things in mind, I wanted to just point you to the last few pages of your discussion guide. Jen so graciously sent us the table that she had created for our own purposes and we kind of have reshared her work here at the end.

And basically, what she did is created a table so you could kind of see all of the measures together and where they were kind of similarities and differences across the few different aspects of the measures. And so, I would just encourage you kind of, as we make our way through take a look at that. You could use it as guidance for some of the key testing results in terms of case size and so forth.

I will point out that the rows that have grayed out are the measures that we will not be discussing. So, you can focus on the (unintelligible) within that table, that will be the focus of discussion.

At the top, she did also kind of pull out some of the kind of overarching methodologies that kind of crossed the all - (eight) of the measures that will also hopefully be helpful in some of this discussion.

I wanted to just also kind of …

Jack Needleman: This is on pages? Ashlie, this is on pages …

Ashlie Wilbon: Sorry.

Jack Needleman: … 24 to 26 of the discussion guide?

Ashlie Wilbon: Yes. It starts on Page 24.

Jack Needleman: Great. Thank you.

Ashlie Wilbon: Sorry. Thank you for pointing that out, Jack. So, I did also just want to kind of as a point of process that we - to the extent that we can be consistent across, so when we make a decision a particular threshold that we're comfortable with, with reliability, for example, that we are consistent with that. I will do my best to point those - point the discussion out and make sure that we're carrying those across all of the measures and we're possible maybe we can either find some efficiencies with getting through these three measures.

So, thank you very much, Jen, for that. We appreciate that. And let's go ahead and dive in.

So, we're going to go to 3508 which is the PCI - outpatient - Elective Outpatient PCI. Sorry, I'm just kind of organizing my papers here. So, again, I won't rehash all of the specs, but this is - a cost measure is a new measure.

It's an episode-based cost measure (certified) at the clinician and clinician group or practice level.

Some of the - actually, what I think I'm going to do, I'm trying to figure out the best way to do this. I think what I'm going to do is kind of refer to Jen some of Jen's summary comments and just talk about how these measures are constructed and then we can maybe dive in the 3508. Does that work for folks?

Again, just - if there are other suggestions on how best to do that, I'm certainly open to that. But maybe that's a good place to start. Does that work for folks?

Ron Walters: Yes.

Ashlie Wilbon: Okay. So, these measures, basically, are episode-based cost measures that use PDC and HCPC codes from Part B claims that trigger an acute condition or procedure episode. Services are assigned from both Part A and Part B using a higher full set of assignment rules.

Both of the episodes are 30 days long so that may vary by the clinical condition. And some of the subcategories -- like for example the revascularization measure -- uses subcategories or stratas to categorize various groups of folks within the clinical population.

And so, the cost measure is it being - is the sum of the ratio (unintelligible) expected payment standardized cost for all the cases that are attributed at the 10th level which basically kind of the group level board, the 10th NPI which would be the clinician level. That sum is then multiplied by the national average of their cost to generate dollar amount. Okay.

So, I think maybe - so, further rules, basically, for seizure episodes are triggered by the presence of relevant service code on a provider bill. Cases are limited for the relevant DRGs so they're trying to assign claims.

Acute episodes are specific to the (integral) clinical condition and the activities of the clinician. And the pneumonia episode is slightly different and that episode is triggered by the presence of a pneumonia DRG on the Part B provider bill.

Let us skip down just a little bit here to risk adjustment. So, the risk model is based on HCC model with 120-days lookback period. And there are also some patient demographic factors that are adjust like needs or disability without ESRD.

Most - in terms of how they handle outlier, a (loud amount) was derived at the point fit for (10th file) and then renormalized by multiplying each episode, was derived expected cost by the subgroup's average expected cost. And then once the outliers are (booted) then the distribution.

So, again, just a high-level overview of some of the kind of underpinnings of the construction of the measures and, again, I'm going to hop back up to 3508 and we will …

Jenifer Perloff: Maybe just to pause for a second while you're reorienting, any colleague should correct anything I said that may have been -- not correct in that summary. I did it kind of quickly across the set. So, if I got anything that’s actually incorrect, please don't hesitate.

For example, I'm looking at this now and not remembering whether 30 days with most of them or not. But again, there were a lot of them. So, anyway.

Ashlie Wilbon: Yes.

Jenifer Perloff: Any clarifications from colleagues are appreciated.

Ashlie Wilbon: Thanks, Jen. Well, we can sort out some of that with the individual review.

Jenifer Perloff: Okay. Great.

Ashlie Wilbon: We'll move back …

Jenifer Perloff: Yes. That's fair enough.

(Crosstalk)

Ashlie Wilbon: So, yes, I just wanted to provide a high-level. So, with the PCI measure, it's looking at evaluating some measure's risk-adjusted cost for Medicare beneficiaries who received an elected PCI in an outpatient basis. The cost measure score is the clinician's average risk-adjusted cost for the episode group, average (above) all the episodes distributed to the clinician.

Again, the risk adjustment uses higher (HBCs). There was no social (background) included in the risk adjustment based on the empirical analysis that they completed and I will - it just comes down to reliability now. This was a sub criterion where there was consensus not reached. We had zero high, three moderate, three low and zero insufficient. We had a split right down the middle.

I wonder - Okay. So, we had both measure score and information on data elements reliability submitted. However, the data submitted or the information submitted for data elements reliability doesn't meet NQF criteria for demonstrating reliabilities of data elements.

So, the analysis in voting for these criteria should be focused on the other empirical analysis that was submitted around the signal-to-noise and test and retest and so forth.

So, for the measure score reliability, again, they did test/retest with correlations and signal-to-noise. For the test/retest, they conducted it using two sets of episodes assessing the correlation and quintile (rise) stability between the 10th - for the 10th NPI (half) scores and then - that was calculated from both the samples and may rank clinicians by their score within each sample into quintiles and then casually percentage of clinicians who changed in the measure score quintile between the two samples.

So, with the signal-to-noise analysis, they had a mean reliability of 0.726 at the 10th level and 0.531 at the 10th NPI level. For the test/retest group in correlation, they had a score of 0.48 at the 10th level and 0.42 at the 10th NPI level.

And zero point work you and you and were largely around the low of reliability for affiliate NPI level and below poorly for the result is that some issues around peace and whether or not they were clearer were enough in

And on (unintelligible), we're largely around the low of reliability score particularly at the 10th NPI level and below treatment correlation 4.

There was also some issues around the specifications and whether or not they were clear or specific enough in order to kind of recreate this measure. And also some concerns around whether the measure is actually recognizing cost for only one provider versus the old PinnacleCare team and also how zero-dollar claims are handled, it was not clear in the specification.

So, with that, I'm going to open it up to the panel for discussion. We do also have a developer from Acumen here on the phone. And so, as questions arise, we will make sure that they have an opportunity to respond.

Ron Walters: So, this is Ron. I have a question that perhaps is at least my difference between moderate and low. So, I think you gave the exact numbers that we all had and are in the analysis that we did or our ratings that we gave.

So, is there formal definitions of what the means need to be for signal-to-noise and the test/retest Pearson correlation score so that we can differentiate between moderate and low? Because once the testing done, absolutely. Are the results exactly as you said absolutely?

The tough part from an analytic perspective at least for me was if the results were adequate or not and whether we had backing to say that. So, that's the first question I'd like to ask.

Ashlie Wilbon: Right.

Jack Needleman: And this is …

Man: Ron, this is ….

Jack Needleman: … Jack and …

Ashlie Wilbon: I was on mute. I was on mute. Sorry. I was talking. I just wanted to say real quick that NQF does not provide any threshold for reliability. There's varying forces of literature that suggest certain cut-offs.

But we don't necessarily have any specific guidance on that and that's why we have convene to you guys. I think there's just need to be some agreement of - about your comfort level for certain levels of scores.

Ron Walters: Thank you for confirming what I believed was true but did not know if I didn't know something that everybody else knew. It is somewhat subjective. Okay. Thank you.

Sri Nagavarapu: And this is Sri Nagavarapu from Acumen. I just wanted to note very quickly that we have reliability numbers for higher case minimum as well that could be relevant to the discussion. We're happy to walk through those.

Those reliability numbers are substantially higher in higher case minimums and my understanding is that that CMS would consider higher case minimums for any of these measures. We presented one case minimum in the initial submission but are happy to present the results at other case minimum which looked substantially higher even than the (unintelligible).

Jenifer Perloff: So, this is Jen …

Jack Needleman: Hi, this is Jack Needleman. I meant to get myself unmute, managed to shut my phone down and get off the poll and I'm back. Where are we in the discussion because I'm really (waiting) here?

Ashlie Wilbon: Go ahead and jump in, Jack. We're still - this is Ashlie. We're still just talking about the signal-to-noise analysis and the test/retest analysis.

Jack Needleman: Terrific. Okay. So, the first thing I want to do is I want to thank the developers here because this is - partly, it's a well - having sat on the cost committee for a while, this is a well-established set of methodology. They're well described. The variations from measure to measure are usually well described although the attribution particularly in the multi-member attribution is always - is a little fuzzy.

The risk assessment models relatively standard with - and they've indicated where they've added some measure-specific software strata. So, I'm very happy with the way in which this information is presented.

And one of the reasons why I'm happy about the way this information is presented is the way in which they presented the reliability data. They've given us something we haven't seen in other measures before which is we see the reliability measures, the signal-to-noise. We see it as a test/retest and we see in the quintile analysis the sense of how stable the rankings are of individuals within these distributions which can be roughly correlated with some sense of the Pearson correlation and the reliability.

Ron was absolutely right. We don't have standards for where we are. So, each of us is sort of applying our own standards and I would love for the steering committee to find some time - not steering committee, the full committee to find some time to actually discuss given results like this where we can see the (input text) what the standards would be.

I think the literature-based classifications are wrong. And 0.4 on Pearson correlation is simply too low for - to be confident that from sample to sample, individuals are being well or being consistently and accurately ranked.

And that was my reaction to this one. If you go back to the original Adams paper that was published on signal-to-noise, he says that a 0.7 on reliability level in a simple split of low, not low, 20% of the folks getting this classified. It's 20 or 25%. I've been (lousy) since I wrote the taper.

But that's enough small number but it may be a number we're prepared to live with as saying we're not going to get it perfect. Everybody is not going to wind up being ranked the same way.

But I feel - I start getting feeling very uncomfortable when I see a lot higher than 20%. The test/retest here is very low. My vague recollection again from some piece of literature says test/retest less than 0.8 is probably not appropriate for individual ranking or individual assessment and that's the way these measures are being used.

And I look at the quintile stuff and that's the piece that's often the missing to me, that's the real practical how stable are the rankings. And the (mid thing), you get put into a decile and the number of points you get depends on where in that decile you are.

So, if you move four deciles to three deciles or two deciles, you move three or four or five or six deciles down the distribution because of the sample that we wind up with, you've just lost half of your - you've either gained or lost half of the points you're going to get for this measure. That feels to me too much.

So, when I looked at quintile rankings which say only 40% of the folks stay in the top quintile which means within the two and 30 - a third of the sample is moved down to the bottom three quintiles, that does not feel stable enough for me to endorse as a reliable measure. And that's why I voted low on this.

Jenifer Perloff: This is Jen. I couldn't agree more and the one addition I would make on the test/retest, this is no variation in the timing of the data. So, it's one big piece of data with two random samples pulled.

There is to me also risks of changes in coding practices, all kinds of coding things that happened over time and (unintelligible) would be an upper estimate because of all of the subtle sort of issues with claims over time. So, that bothered me as well.

Christie Teigland: This is Christie Teigland. I just want to weigh in and I totally agree with Jack as well. I think the movement across the quintiles, you can have a score of 95, one period or one sample and then across 65. I mean, that's probably - if you rank looking at some rating system, a five-star compared to a three-star, two-star, I don't know what - it could be a huge difference in where you might rank and that's exactly what we don't want to see.

Ron Walters: So, again, in our - this is Ron again, in our advisory role to the TEP steering committees, thank you, Jack, for that explanation. I think even though you didn't formulate a rule, you certainly explained the rationale for a potential role.

I think that's exactly what the kinds of things and I probably would consider changing my score to low with kind of a synopsis of exactly what Jack said. We need to convey to the TEPs that while there are not formal definitions for moderate and low reliability, the concerns we have when the reliability scores

start to hit these sorts of numbers that the reliability measures under serious threat.

And how we do that is not necessarily so much a matter which box we put it in lower moderate although that's only the mechanism we have available as the overall message communicated that there was a great deal of discussion in the committee about the reliability scores that were presented and great concern over whether this measure would be applicable across different sizes, different groups, different times, all the things you just heard. I'm okay with that.

voltages for on the other three are from ecumenical bed could address some of the common, yet the who wanted to mention that on the noted the reliability result that you presented without your measure were using a case minimum of 10 of those on the codes and offices measure of the

Sri Nagavarapu; This is Sri Nagavarapu from Acumen. I was hoping I could address some of the comments that have come up so far.

Ashlie Wilbon: Yes, please. Go ahead.

Sri Nagavarapu: So, the first thing I wanted to mention is that as I noted, reliability results that we presented for the outpatient PCI measure were using a case minimum of 10 episodes just because in the past, this measure had been considered at 10 episodes.

But this is a choice that CMS will make down the road. And so, the reliability is substantially higher at different case minimum. So, if you move to 20 episodes for the version of the measure, the mean reliability is 0.802. If you move to 30 episodes, it's 0.841 and then 40 episodes is even higher at 0.87.

Similarly for (TMTIs), you get very sharp increases in reliability at different case minimum. When you move from 10 to 20, you move up to mean reliability of 0.65. Going to 30 is at 0.72 and then going to 40 is at 0.77. And at that point, even the minimum reliability is 0.701, above the typical threshold for high reliability.

In regards to the comparison between those reliability numbers and the Pearson correlations for test/retest, the one thing I would want to caution people on is that all of these are done with a 10-episode case minimum. And so, because we are using one here of data for the test/retest and we do a random sample, we're looking only at the set of providers with 20 episodes or more and ensuring that they have at least 10 in each sample.

Now, that's something because of the sharp increases in reliability that I just mentioned that that Pearson correlation is very likely to change if we use a different case threshold than 10 in order to do the correlation across the test/retest. The quintile rank stability is also very (unintelligible) based on case minimum for that reason.

And so, for that reason, the Pearson correlations here should in fact (be thought of as) a lower estimate, a lower bound estimate of what the correlations are because we're starting from an extremely low case minimum.

The final point I just wanted to make was about the overall process of risk adjustment. If it's helpful, what we can do is for the summary that was given for all of the measures provide sort of feedback and review the summary, the measures to correct anything there.

There were some points made about risk adjustment where I just want to emphasize that the risk adjustment models do incorporate measure-specific risk adjusters that were suggested by the specialty society representatives that we worked with closely at each step of the way to construct the measure.

So, the episode windows were - they vary. For the measures you're considering here, they vary anywhere from 30 to 90 days. For other measures, they could be shorter. Those were clinical determinations made by the specialty society representatives as were the addition of other risk adjusters.

I'll stop there but I just wanted to make sure that the numbers on reliability were considered because I know you haven't had a chance to see what they would look like at higher case minimum and they look substantially higher.

Susan White: So, this is Susan. Just on final clarification just so I understand the rules here, so, the measure developer submitted the statistics for 10 and that's the measure that we're - that's the conditions under which we're assessing the measure. Is it okay to - I mean, we haven't seen the numbers, they're going to be higher, I mean, mathematically, they ought to be higher. So, what's the - what are we supposed to do with that information I guess is my question.

Ashlie Wilbon: Yes. That's correct. This is Ashlie from NQF. So, generally, we do not allow at this point in the process developers to provide additional information. So, to the extent that they would be adding in new testing data.

So, the method panel should be voting on what's in front of them at this point and we can work with the developers on how to fix the submission or provide additional reliability results potentially in another submission.

Jack Needleman: Okay. Sorry, this is Jack. Just really quick, in terms of specifications, there's a minimum number of cases for the specification.

Ashlie Wilborn: Yes. I believe it's set as well but we can clarify that right now. Sri?

Sri Nagavarapu: Yes. Our understanding is that in past NQF submissions that we've seen that the case minimum isn't part of the submission. This is CMS decision but I know that they're not wedded to a 10-case minimum on these episode-based measures and there's always a discussion of the trade-offs between reliability and submission coverage in order to make that decision about a case minimum and these sorts of numbers that I'm giving to you which actually we could send you right now for - at the end of the day or however you prefer, these are the - those scores would go into the term with that case minimum at CMS.

Ashlie Wilbon: Yes. The case minimum - this is Ashlie. The case minimum is very similar to the discussion that we had with the (LAP) measure. I mean, we're really (cutting totally) the line here in terms of changing the application which did not or should not be changes, fixing a typo maybe (unintelligible).

Adding information to the submission at this point generally is not part of the process as we don't really have time for folks to be able to reassess new information and still be able to kind of stick to the (plan) of the process. I think sharing that information is helpful but I think at this point for where we are on the call, we need to have folks vote (on this) in front of them.

The case minimum is technically for specifications because, again, like what we said earlier this testing of the measure also is representative of the threshold of how the measure can be used. So, it is technically part of the specification. So, yes, I'll just stop there.

Sri Nagavarapu: Yes. And I just wanted to - I think someone on our team mentioned that maybe the connection wasn't clear but I just want to make clear that CMS would - my understanding is that CMS would be willing to consider higher case minimum in order to consider the reliability trade-off. I just wanted to make sure we're clear with this one. Thanks.

Ashlie Wilbon: Okay. Thank you, Sri.

Jack Needleman: Great. Ashlie, just - I was just struck by your language. Technically, the case minimum is part of the specification. But I would go one step further given that what's - what we know, what just got explicitly said, what's documented in some of the numbers is reliability is very much a measure of - the individual was very much a measure - matter of the caseload and the size of the number of cases.

So, we can only consider making a judgment reliability at a specific level of sampling or cases. And this is one of those cases where there's real difference between us and CMS. CMS needs to be - CMS worry - has to worry about coverage. If we make cases so high, we've got four people in the country covered by the measure, that's not - from their perspective, that's not useful.

So, they got to think about the reliability. They may be thinking one way about reliability case, the number of folks covered trade-offs. From our perspective though, there's a minimum reliability threshold we need to be comfortable with and that trade-off is different in the calculation of CMS than it is for us.

The trade-off with how many people the measure will actually affect is irrelevant to the judgment of does the measure meet the minimum level of reliability we're comfortable with.

Ashlie Wilbon: So, are there other comments method panel members on this particular issue and do you feel ready to go on reliability at this point for this measure?

Ron Walters: I am. This is Ron.

Ashlie Wilbon: Yes, no from others?

Susan White: It's Susan. I am.

Karen Joynt Maddox: Yes. This is Karen. Me, too.

Christie Teigland: Christie. Yes.

Ashlie Wilbon: Okay.

Ron Walters: I am yes.

Ashlie Wilbon: Okay. Thank you, everyone. It's hard on the phone to gauge where folks are. So, I'm going to go ahead and ask you to submit your votes for reliability via the SurveyMonkey.

Again, the ratings for validity were moderate but I did just want to brave one question for the committee on moderate - on the validity criteria in which would be applicable to all of the measures. The face validity assessment that was submitted does not meet NQF requirements. So, we do require a systematic assessment of face validity so there has to be some description of a process where an external group - a group external to the measure development process was asked if there - a question or series of questions

about the ability of the measure or the validity of the measure either via scoring or survey or some sort.

And so their description of face validity doesn't quite - does not meet that threshold. They did submit empirical validity testing for each of the measures. And so, for this particular measure as the case, I just want to make sure that folks are voting - we're voting not necessarily a moderate based on accepting the face validity but that the empirical validity testing actually - that your assessment of the empirical validity testing actually was acceptable.

And in this case, they did concept validity with indicators that were shown were identified based on being any series of resource utilization, would go around hospitalizations and post-acute care. This was similar to how they constructed the empirical analysis of validity for - of all the measures or for this particular measure.

They used mean (observed) expected cost ratios for various indicators around whether or not it a hospital admission or whether or not there was a pack - post-acute care (unintelligible).

Sri Nagavarapu; Okay. And this is Sri from Acumen. On face validity area, I was wondering if I could just make one comment to describe the process that we sort of walked through and the testing.

Ashlie Wilbon: Sure.

Sri Nagavarapu: So, essentially, we have a comprehensive process of collaborative relationships with specialty societies that are relevant to each of these measures. We go through a recruiting process where a large group of

specialty society members decide which episode-based measure to create and they vote based on threshold.

We then moved to building each measure in detail with clinical workgroups that are pulled from those specialty society representatives. They vote at each stage of the way on measure specification on each aspect of trigger codes that define the measures, the risk adjusters (that could present) our costs they counted.

Typically, we use 60% threshold to have a (formal loading) process so that we can document whether decisions were made by the work group. After that, we went through a field testing period for a month where we released report to clinicians nationally. To my knowledge, it's the largest field testing effort that CMS has had for any measures.

This year, there are about 800,000 reports released to all distributed clinicians naturally across all the measures. And based on the feedback from that field testing period, we go back to the clinical work group from the specially societies and walk through the comments that we got in filed testing. They look at measure specifications and make changes and then vote on changes to measure specifications.

So, I just wanted to - I know that we walked through the clinical subcommittee and the work group process in the face validity section of the testing form and I wanted to make sure that it was clear that the formal process with voting at each step of the way on measure specification in sort of comprehensive engagement with the specialty society was directly (included).

Ashlie Wilbon: Right. And - but - so, I think that clarifying point is - there's a couple of things with that is voting as a part of the measure development process as

opposed to some external assessments of the measure's validity I think is kind of what the distinction would be.

And so, maybe if you could clarify those voting, was that part of the development process? And so, the people who were voting were also helping to establish specifications and kind of structure the measure.

Sri Nagavarapu: Yes. The work groups were those who are structuring the - they work on developing the measure specification. The voting process though I would describe as an external input process in the sense that all of these work group members were nominated by specialty society to join this and the work group members have the opportunity to discuss measure specifications with their specialty societies both (test work) and pure field testing.

And so, the way that process is set up as a way of getting external validation of the measures that at step of the way, just leveraging the fact that the work (steps) in the subcommittee members have already (seen) the details of the data and be able to communicate the questions to specialty societies and others they're affiliated with or representing (them).

Ashlie Wilbon: Okay. And I don't want to play with this too much but I think the other piece that we would be missing for that in order for us to - at NQF to kind of consider that face validity is we have to have some understanding of how it would systematically assess.

So, everyone kind of ask the same question or set of questions about the validity of the measure, the third point in the process, kind of external evaluators of the measure to say that they could get some sort of kind of systematic assessment either by ranking or voting or scoring so that those results could then be collated and evaluated based on the percent of folks that

scored nor that they strongly agree that the measure was valid or what have you.

So, I think again when we take systematic assessment of face validity, we're looking for not folks who are able to get feedback but what was the kind of structure around them saying that this was a valid indicator of how the measuring is doing, what is intended to do.

So, yes, I'll just leave it at that and see if other method panel members have anything to add to that.

Sri Nagavarapu': Yes. And there's a vote after field testing clarified to ensure that people were satisfied with the measure specifications and based on the external input that was given through the field testing specialty society. So, I just want to make sure of that.

Ashlie Wilbon: But those votes were not included in the submission, correct?

Sri Nagavarapu: Those votes aren't in the testing submission. The votes were taken along - throughout the process as well as after field testing on every aspect of this (area) but they weren't included in the testing submission form.

Ashlie Wilbon: Right. Okay.

Sri Nagavarapu: Okay.

Ashlie Wilbon: Okay. I think that's helpful and I do think it's probably not quite there based on our requirements. So, I think that it sounds like you guys might have done something like that but maybe that's not quite communicated in the form in

the way that we need to kind of have it reflected in order to have it kind of meet that standard.

And so, I think the question is really then around the method panel's evaluation of the empirical analysis that was submitted. And so, I did want to just give them an opportunity to request on that and to see whether or not there were any concerns with that.

And if there aren't any concern, then we can - you don't need to rehash validity again. I just wanted to kind of make that distinction and just make sure that folks have an understanding and that the votes were kind of based on the same assumptions.

Jenifer Perloff: This is Jen. I absolutely did not pick up on that sort of face validity distinction. And so, my ranking absolutely considered face validity. Anyway, I just wanted to - this has been a clarifying conversation for me.

The other question I would ask to my colleagues, I think someone pointed out earlier that there's a pattern across all of these measures and there's sort of hypothesis testing around who should cost more. It's interesting how the statistics behave depending on the measure. This one looks somewhat well behaved where the sort of residual group is close to one and the high cost group has a ratio of 1.48 or 1.07.

Some of them, the variable that they selected to split the group on almost look like a perfect predictor of the score. And so, does this - the behavior of the statistics was curious to me and, I don't know, I was curious what my more statistically-trained colleagues would have to say about this. Is that okay to ask a question like that? I'm sorry if it's not the right venue.

Ashlie Wilbon: No. I think that's fine, Jen. Definitely.

Jack Needleman: Okay. So, this is Jack. Those are useful and, yes, because in some cases what they've done is they picked the problem - in some senses, it's - again, we know on some of the cost measures where everything is standardized pricing, once you've included the DRG and its standardized price and sort of the base cost adjustments, all the additional costs, the variations in costs come from the post-acute care heavily influenced by institutional care and readmission.

So, it should be that if you got a readmission, the cost associated with that readmission is going to be higher than - it's going to be what's driving a ratio higher than expected since the readmissions are not predicted or aren't fully predicted by the risk adjustment. If the readmissions were fully predicted by the risk adjustment, then we have no basis for unplanned readmission here or on affected readmission.

Some patients are just going to get readmitted but given that risk adjuster in this one is virtually doing no work at all …

Jenifer Perloff: Yes. Right.

Jack Needleman: … but in our SNF it's doing 20 - it's predicting 25% of the variance or 50% of the variance. I think saying - yes, it's a given that if readmissions are the things that drive up the costs and we haven't fully predicted readmissions, we're going to see a higher-than-expected readmission number. That's what should be happening.

All I'm saying is our excess costs are higher than expected unless it's coming from the readmission or coming from the SNF care, a substantial portion of

them are. So, yes, it's just confirming what we know from the construction of the measure.

Jenifer Perloff: Exactly. So, to me, that analysis didn't really comment on validity. It was really just sort of demonstrating how the measure works. And therefore, my rating was primarily face validity which is now off the table. So, I'm a little bit of a lost - at a loss but.

Sri Nagavarapu: In terms of complementary analyses for empirical validity, other items that we've looked at and we can provide upon request, looked at various types of kind of themes or categories of clinical outcomes on the admissions and post-acute care.

My colleague (Alex Zand) who's a cardiologist here can walk through some of those categories for about patient PCI measure very quickly.

(Alex Zand): Thanks, everybody. So, with outpatient PCI, we divided the cost in the multiple themes, including the cost of the initial PCI, the cost of a recurrent PCI after that initial PCI, the cost of bleeding complications, the cost of other complications unrelated to that bleeding and the cost of other admission and do not necessarily going to the details as we expect that higher costs related to the initial PCI were significantly higher and higher cost quintile providers because there's other providers with higher cost scores had had higher cost there, which (unintelligible) was expected.

But what we also thought was that those providers at higher costs when it came to the recurrent MI or recurrent PCI after the initial one which again would be assigned of these different assets of quality were likely correlated. And similarly, we're more likely to have higher cost when it came to bleeding complications as well.

So, I think that there's some correlation across these different clinical themes that we know are related to kind of the quality of care and the efficiency of care provided.

Jenifer Perloff: I happen to think it's nearly impossible to validate claims-based cost with claims. I think a real validation test requires external data, right? Otherwise, I mean, you can propose hypotheses and tests, I mean, I guess that's get you a certain distance down the road.

But just, again, from an overarching methodological (process) which is very challenging because it's using the same data that the measures based on to validate it. So, I appreciate that. I like the clinical thinking that helps a lot. But just to figure a point.

Ashlie Wilbon: So, we're at about 12 minutes to the hour and I just wanted to kind of do a post-check here, are there others besides Jen who their initial rating for validity was based on face validity? Yes? No?

Christie Teigland: Yes. Mine were, too. Christie. I didn't realize that.

Ashlie Wilbon: Okay. So, there's also kind of a high vote here so it's hard to tell because we're doing face validity, moderate should be - would be the highest vote that you could submit based from the way that the criteria set up. So, I'd like to suggest if folks are okay with this that we do another vote for validity unless there's other discussions that you'd like to have based on the empirical analysis that's here. Does that resonate with folks?

Woman: Sure.

Ashlie Wilbon: Are you ready to vote?

Woman; Yes.

Ashlie Wilbon: Okay. Let's have you go ahead and submit your vote on validity for this measure and you should be able to submit that. So, there will be two votes for this measure. One on reliability and one on validity.

And so, now that that's done, okay, we have 10 minutes left and I think what would be useful at this point is to just have a discussion about how to address or how we should approach the evaluation of the remaining measures. But we do have a second call that is scheduled on next Tuesday which would be a continuation of this call.

What are your feelings about - of kind of evaluating the remainder of the measures that are currently plated for discussion? Should we - are there particular kind of speaking points that you think might be applicable to all of those that we might maybe restructure that discussion on the next call on a different way or should we - are you okay with us just continuing to go through measure by measure and have that discussion as Jen did mention that some of the measures you kind of behave a little bit differently based on kind of the clinical population and the data set? So, I just want to kind of get a sense from folks on what you're leaning is to - towards.

Jenifer Perloff: Sorry for the quiet. I'm just reading down ahead to the ones that are still here.

Ashlie Wilbon: Yes. So, we have the colonoscopy surveillance screening measure and the knee arthroplasty measure.

Jenifer Perloff: Yes. I mean, I think all of the remaining ones sort of had - it's interesting how these all fill out. Right. They have points of kind of some of the empirical concerns that we've talked about, they're right in the middle of being stronger than the weak ones. But they each kind of have their own features.

Ashlie Wilbon: Okay. While we do this then, we will just pick up on the next call and jump into the last two measures. And at that point also we will offer the opportunity again on whether you'd like to pull additional measures for discussion or whether you think any of these overarching issues with impacts any of the other measures.

So, we will - I don't think it's worth kind of diving in in the next seven minutes unless you guys would like to do so. I think we'll still have to see you carry over to the next call and it seems like it would be easy to kind of start fresh with a new measure if folks are okay with that.

Jenifer Perloff: On a Friday afternoon for sure.

Ashlie Wilbon: Okay.

Jack Needleman: Yes. I'm okay with dealing with the last two one on the next call. I'll be in China when we do this. So, it's going to be 10 o'clock at night. Let's see if we can get it done a little bit faster.

Ron Walters: This is Ron. I agree.

Jack Needleman: I do want to - since we - I do want to just provide some additional feedback to Acumen and, again, I have been very appreciative of both the level of documentation and the thoughtfulness with which the measures are

constructed and the discussion of the expert panel process was - helped reaffirming of that.

The one comment I would make, if somebody is trying to assess reliability and better understanding your expert panel thing is - and (unintelligible) you're not reporting each individual vote of the group that they've hit the 60% threshold, how high it was, that's going to be an awful lot of votes to report and the sort of general standard sounds reasonable.

But I am - as I read that description, I heard you describe it, I'm not super comparable with the somewhat in agreement category being included within 60%. That's a very weak endorsement level. So, it's your process but I would feel more comfortable if the scale was a little bit different and there was a stronger level of endorsement and not - and I guess I'm okay with it as being included within that 60%.

So, that's just some feedback. It's your process. We'll see whatever gets delivered.

Christie Teigland: It's Christie. I second that. I noted that in my comments as well. I don't think that's be included in, yes, high reliability and face validity.

Sri Nagavarapu: This is Sri from Acumen. I'll just clarify the - and I appreciate those comments. Can I just clarify what's meant by that somewhat in agreement point? Just to make sure.

When we would take votes on particular aspect of the measures, they're the yes/no for let's say should this cohort be excluded or like locked in status for the stroke measure for instance. And so, those were yes/no votes and so I just want to make sure I (unintelligible).

Ashlie Wilbon: So, this is Ashlie. I don't want to answer for Jack or the other method panel members who had a concern about that. But for a demonstration of face validity, it wouldn't be on pieces of the measure. It would need to be a systematic assessment of the measure as a whole and whether or not it is measuring what you're intended to be measuring.

So, I think that's also kind of the distinction there with kind of taking votes throughout the process on pieces of the measure versus having an overall kind of objective assessment of the measure as a whole and how it's performing and what is measuring as opposed to kind of a particular specification or something like that.

So, I just want to clarify that in terms of what would - what we would potentially need in future submission for - to demonstrate face validity. Hopefully that's clear. But I would kind of throw that back to Jack on his comment about the somewhat.

Jack Needleman: Yes. You got - you caught it, Ashlie.

Ashlie Wilbon: Okay. So, with that, I would go ahead and log off. Thank you everyone for your time. We will reconvene next Tuesday and we will pick up where we left off for the last two measures.

Ron Walters: Thank you.

Ashlie Wilbon: Okay. Thank you, everyone. Bye. Have a good weekend.

Sri Nagavarapu: Thank you, Ashlie.

END