NATIONAL QUALITY FORUM

Moderator: N/A
March 11, 2019
4:17 pm CT

(Bijan): Hi this is (Bijan)

Woman: Hi everyone, welcome to Authentic Methods Panel Approach Meting Subgroup 1. We're happy that you're able to join us today. Just some starting points for our day-to-day and then we'll jump into a roll call and start discussing our measures.

A discussion guide was sent out to the Subgroup member Friday evening. This document will be our guide for today's meeting and for the measures discussion. We will follow the order presented on this document consensus for the first seven measures means that we will be talking about them and that information has been provided on your discussion guide.

All other measures will not be discussed during today's call unless a member of the Subgroup would like to take this opportunity to pull a measure. If you choose not to discuss additional measure, this is a premier play analysis and will be made final. And now just pausing for a second to actually, we will come back to this. We'll do roll call and then we'll ask if anyone would like to pull a measure for discussion for your knowledge.

In the same email that was sent out that had the discussion guide, there was a link to a SurveyMonkey, just to make everything a little bit more efficient we ask that you pull that SurveyMonkey up now and start casting your votes as we goes through all the liability and overlay discussions at the conclusion of each measures discussion.

And you are free to start putting it in and then hit Submit until the end. It just makes it easier for us if we get the results now instead of having to, you know, track you down later. And we will obviously prompt you every five minutes time for you to cast your vote.

Timing is limited on today's call. We have roughly 15 minutes to discuss each measure. We do have a lot of measures pulled for discussion already. And although we would like to come to present all this measure today, we do have a follow-up meeting scheduled for Friday, March 22 from 1:00 to 3:00. That is a backup meeting.

The goal hopefully we're able to get through the discussion of these measures today. But if not, we do have that translated. And then I just wanted to remind everyone that this is a public call. The developer representatives are on the line to answer questions from staff or panel members. However, there will be no opportunity for public comment. For record keeping, we ask that you every time you speak up that you say your name so we know who's speaking and we can keep track of that.

Before we do roll call, I just want to see if there's any question about the structure of today's call?

(Bijan): All right. This is (Bijan), I just tried to open up the link for SurveyMonkey, and it's not opening up for me.

Woman: What are you getting is it just - maybe…

(Bijan): It's just going blank, going to the Web site apparently, but I'm not getting to see anything in the (unintelligible) blank slide.

Woman: I'm going to send the link out again and hopefully that will work for you. Any other questions or concern?

(Bijan): No worry. You know what, this is? I have two computers. The window based one is working. So, the Mac one is not. Oh, that's fine, now I get.

Woman: Okay. Perfect. Thank you. All right. Well, that we will just go ahead and do roll call, May.

(May): Okay. So, hello everyone this is May. I'm just going to do roll call for the assigned method panel members, who are assigned these Subgroups measures. And again, just to clarify, they're the ones who are, who we need to tap final votes and use the SurveyMonkey for.

So I've heard you John, you're on the line. Yes?

(John Bernot): Yes ma'am.

(May): Okay. John Bernot.

(John Bernot): Yes, I'm here.

(May): Okay. (Paul Comaske)?

(Paul Comaske): I'm here.

(May): (Jeff Caper)?

(Jeff Caper): Here.

(May): And (Sherrie Chaplain)?

(Sherrie Chaplain): Here.

(May): All right. Thank you. And then for all the other panel members, just to feel free to listen to it.

Woman: And I just want to remind everyone that these calls, we did change our structure, our conferencing system. So all lines are open but we do ask that everyone mute their line unless they are talking, just to reduce background noise.

Okay. With that I think we're ready to start. And I believe Michael is going through our first measure.

Michael Abrams: Yes, so we're going to start at the top of the discussion guide measure number 3506. This is Michael Abrams talking here NQF. The measure title hospitalization after release with missed emergency department dizzy stroke, and I'm going to try to go through. And in three to five minutes, the discussion guide just as it's laid out here and I'll try to use headings and stuff so you can follow along with me.

This is a new message, brief description. This measure is a measure of rates of patient admissions to the hospital for a stroke incurs within 30 days of being treated and released from an emergency department for symptoms that were presumed at that time during that emergency department visits to be benign dizziness or inner ear vestibular diagnosis.

The numerator and denominator of this measure both are predicated upon primary diagnosis. Index visits for the denominator are the first one that appears in the window or the period window of performance. After that index visit, a person is not reentered into the analysis for 360 days following that.

This is what would be referred to as an outcome measure. In other words, it looks at in particular stroke, a stroke event leading to a hospitalization is the outcome if the claim is based measure. And testing was done by the developer using for Johns Hopkins Hospital Emergency Department venues 2152 emergency department visits were evaluated as part of their evaluation of their scientific testing.

And then the Chris or the Chesapeake Regional Information System dataset, which has data across the 47 regulated hospitals in Maryland was used to supplement the data. The data that's what that spans 2012 to 2017.

There were no specific exclusions listed in the specifications for this measure, presumably because as rather the specifications are specific enough to identify the cases and the venues. The level of analysis we're talking about here is facility hospital, hospital EDs.

Risk adjustment was in fact done and tested and that was based on looking at 91, the 360 days' post. So three months to 12 months post the emergency department visits to find if you will baseline rates of stroke events happening.

And then using that to adjust for the 30-day follow-up stroke, which is the focus of this measure.

The reliability testing will move right on to that. And there's that specific bullets. And I'm right on approaching the bottom of this first page. For the liability testing, you all three of you rated as moderate to as low. That's why we're discussing it today. Consensus has not been reached.

Let me remind you what that reliability looks like. The developers provided estimates of stability within hospital over three successive time periods of roughly 18 months each and then gave 95% confidence intervals in their discourse.

They say that those confidence of windows were generally stable with time even as they did vary a little bit, particularly for one hospital but they were stable enough over time to suggest to reproduce ability or stability of the measure within hospitals.

So that was the description of reliability. You know, perhaps it's a good opportunity to pause for a second and see if the committee members wanted to clarify anything about reliability in discussion or think about their boat that they made and if they want to revive, they're okay with that.

(John Bernot): Well, this is John Bernot, may be not specifically reliability. But in the summer, you know, what you just noted that this was risk adjusted, but the measure Stewart in submitting this in response to a question 2B 3.1, they say no risk adjustment or stratification, so there's some discrepancy there.

Man: Discrepancy okay.

(Bijan): Yes, that's my, this is (Bijan) from the opening. This is, that's my point too. I'm just kind of through that and I think as John pointed out, I think I had the same issue.

Man: I have a question…

(Paul Comaske): This is (Paul Comaske), I think it's because they did a very tricky way of doing. I say that was very smart the way they did risk adjustment. They say will it be the expected baseline rate for stroke over the next 30 days in this specific population that showed up. And then the answer to this question, do we - showing the risk hazard stroke for risk hazard function for stroke after he visits in that population.

So they therefore got the 90-day, the 90- to 360-day period calculated but it was, they use the same population. So it actually risk adjusted itself. So in a way they didn't do an outside model risk adjustment but they did account for risk within it.

Man: Yes, that's correct. That's my reading of the application as well that they within individual did a risk adjustment based on their, the follow up is three months to 12-month period posts the ED admissions.

And any other comments or concerns or perhaps one of the one of the key comments is that the developer should in the future clarify specifically that they applied this risk adjustment to the scientific testing? It's my reading that indeed they did, but there sounds like there's some confusion among other reviewers about that point.

(David Newman-Toker): This is (David). I just wanted to let you know that I'm on the phone, if you have questions you'd like to answer, I'm happy to do so.

(Karen): Thank you (David). So this is (Karen) from NQF. I think it'll really help us, I know we've already kind of JumpShift and we were talking about risk adjustment. Let's try to do this in order of reliability first and specs and go in order. I think that'll just be a little easier for us to make notes and as to make sure that everybody has voted in the way they need to vote.

So it sounds like we mostly after risk adjustment, we'll come back and tag that one in a minute. But let's just go back to specifications, and even before that, do we need to do, I think we need to, this is apologies on my part. We actually need to do some disclosures of interest. So, those of you who have been with us before, remember that we do this on occasion.

So I think we need, what we're trying to do is just ask our methods panel members to disclose anything that might that, you know, things that you do in your other non-QF life that may have a bearing on any of this work. And that could be paid work or voluntary work, that sort of thing. And things that you disclose are not necessarily conflicts that we just want to make sure everything is disclosed.

So if you don't mind, let's go ahead and I'll just go through the list I have May did it. So (Bijan), do you have anything that you want to disclose?

(Bijan): Not specifically but I do have some projects, something completely unrelated with John Hopkins OB/GYN department.

(Karen): Okay. But nothing that has to do with any of the measures under discussion today?

(Bijan): No.

(Karen): Thank you, John?

(John Bernot): Yes, nothing to disclose in relation to these measures. Thanks.

(Karen): Thanks John. (Paul)?

(Paul Comaske): Nothing to disclose.

(Karen): Thank you. (Jeff)?

(Jeff): Just in general that the telehealth, the measures management contract with CMS, which manages portfolio measures for accountability programs but other than that nothing to disclose.

(Karen): Thank you (Jeff). And (Sherrie)?

(Sherrie): I'm on the NQF All-Cause Readmissions Committee. I sit on our hospitals re admissions committee, and I'm on the Yale Core team, MIT inpatient measures, Technical Advisory Panel. But none of them specific to the measures I think I have today, because I think I would fill that out in advance of this whole review process.

(Karen): Yes, we asked you before each cycle or as we know which measures are coming through and then we have you publicly state on our calls. So thank you so much.

Okay. So, and sorry to interrupt the flow of your discussion so Michael, reliability, how do you do you want to do that, do you want to start with

specification and then go into the testing and what was or wasn't provided and then kind of walk away through?

Michael Abrams: Yes. So we were talking about reliability, you want to start with specifications and discuss that initially. So are there any, so the way I would summarize, the points I think up in the top of the page summarize specifications pretty well with regard to primary diagnosis being identified both in the numerator and denominator.

The only comments that came up from you all for discussion is you'll find that the bottom of the page of the second page under items to be discussed is the emergency department venue well enough to find was brought up by one individual, given that these are all facilities and regulated hospitals. Perhaps that isn't a point that needs much discussion. But there was some question about that.

And another with regard to the specifications specifically is our hospitalizations wall enough defined as well with regard to facility type and the observation of the stay. Was there any concern about that?

Again, given the venues and the data that was employed, that may not be a lengthy discussion but were there any concerns about how the measure was specified specifically to isolate the count of cases and then of cases that achieved this measure of seeing a miss stroke event, anything from the committee about that?

(John Bernot): This is John Bernot. I think those are probably my observation. I mean, for the sake of Hopkins doing this with Hopkins facilities, yes, that's fine. That's they know what their emergency departments are. But I mean, the idea of is

this measure adequately specified would be implemented in a standardized way.

User need to specify what constitutes an emergency department visits, what constitutes a hospitalization. And unless I missed it, looks pretty thoroughly and that specified section, they don't define what constitutes an emergency department as a Nora hospitalization.

I mean, of course, in that, like, for example, in the field of readmissions, it's much discussed. This is an observation stay hospitalization and, you know, it is a minimum of x hours or expertise to constitute hospitalization, what types of facilities is called the hospitalization? Does that, you know, span long-term care facilities, is the acute care facilities only et cetera, et cetera. These need to be specified if the measures can be implemented in a standardized way.

Michael Abrams: Okay. Any other discussion about that from other committee members?

(Sherrie): Not about that but I have another issues.

Michael Abrams: Okay. Another issue, is it related to specifications in particular, just so we keep on focused on that for the moment?

(Sherrie): This is Sherrie. I found this, the specification daunting. I thought they were pretty, they're pretty complicated. And at a university teaching hospital, I get it, rolled out into the field. I'm not so sure that these things can be implemented as well. But I found three things confusing.

One, the sampling with replacement. If the patient is replaced in the sample post be discharged, you know, after one year but before three years is, I got

lost about with the, does the patient go back into the sampling pool after one year but before three years is developer on the line can they clarify that?

(David Newman-Toker): Yes, it's (David). Sorry for the confusion. That's exactly correct. There's a one-year follow-up period (unintelligible) because it's darn rich...

(Bijan): Can you speak up, I can't hear.

Michael Abrams: Yes, you're very difficult to hear.

(David Newman-Toker): My apologies. Let me, it there better now?

Michael Abrams: Only slightly.

(David Newman-Toker): All right. Hang on.

Woman: Hello?

(David Newman-Toker): Can you hear me better now?

(Bijan): Yes.

Michael Abrams: Yes, much better.

(David Newman-Toker): So sorry about that I apologize I'm on a train. So there may be some ambient noise but yes, the reviewer had the correct assessment which is that we took the approach of saying that you get one bite at the apple a patient as an ED index visit. Then for the rest of that year, you're only looking at the outcomes for that patient.

But then you can be renewed and have another ED dismissed visit and a year later, as long as you're still within the assessment window, the performance period. And again, we're happy to, you know, if there are some specific suggestions for how we could enhance the measure, we're happy to consider those. We appreciate your help.

(Sherrie): I just think that that needs a lot more clarification because if I were, you know, if I were monitoring following these patients, presuming they are candidates for readmission within that time window it's pretty, the way it's written now, you know, it took me a while. And this is what I do, to kind of figure out who was actually in the sample and who was excluded. So that some more clarification in more detail I think would be very helpful.

(David Newman-Toker): We're happy to do that. Thank you.

Michael Abrams: So any other comments about specification specifically right now? If not, we'll move on and talk about reliability directly.

(John Bernot): Yes, this is John Bernot again, their response to S5 and the MIS form, there's a very up to sentence that states additional data sources may be required to fully understand admissions that occur out-of-network, the missing data can be approximated.

There's nothing that talks about the methodology for approximation that's acceptable, how is that out-of-network defined et cetera, et cetera. That this is just an ill-defined concept that again, administer the measure uniformly across number of entities, you know, that that that will not cut it.

Michael Abrams: Yes, agreed. And we'll talk about that at the, when we get into validity, there's a 17% loss that's expressed in the application related to out-of-network

and clarification of that. It was one of the things that you all were concerned about. So we will talk about that and validity. Anything else specifically about specifications before we look at reliability a little bit more detail here? Okay.

(Karen): Sherrie, this is (Karen), I thought you said you had three questions. Did I misunderstand you or have we lost two of them?

(Sherrie): Never mind. I got lost in.

(Paul Comaske): This is (Paul). Before we move on out of reliability, I'm just a little concerned about the white confidence intervals because it's going to impair the ability to tell to distinguish one place from another place. And if you looked at the, you know, if you look at the values from one place to the next they vary, but if you looked at the conference intervals, they did not distinguish one from the other.

Michael Abrams: Yes, agreed. For the most part, that was the case. So, and that was the key concern about the way that the signal and noise was presented. And the suggestion that this indicated stability, even if there are white confidence intervals and some change even within hospital with time.

So any additional comments about that or thoughts about revising your ratings on reliability at this point or are comfortable with where you're at with three moderate ratings and too low ratings.

(Karen): So this is (Karen). Just to go back a little bit and this is before Michael's time, so he might not have realized this. But we've talked in the past about what score level reliability methodology is. We might be happy to take obviously, signal to noise is one, the split half liability methodology is a second one.

And there's been a little bit of non-clarity as to whether showing confidence interval in the way that this was done would actually reflect a liability.

NQF in the past has accepted those if the confidence stand don't overlap but I think the methods panel has been a little bit more reluctant to accept that. So that's the background a little bit. Along with that background though is also a reminder that if appropriate data element validity has been conducted and the values the results are adequate, then we don't actually at this point, require additional reliability.

So kind of two things for you to think about. Number 1, you know, will you accept what was given as reliability. And if you don't, then we need to think about the data element validation that was done and whether or not that is acceptable because that can be used in lieu of what you don't like about the disability analysis that this developer fit.

(Bijan): This is (Bijan) from (unintelligible) so in terms of data element of valid reliability I mean, the very fact that this measure is going to use claims data under confused as to what exactly are we talking about data element validity or reliability here? Because these are all sort of, you know, at least as far as I understand, I mean they're all sort of already reliable, right? The measures that will be potentially be used in this major?

Michael Abrams: Right.

(Karen): Yes, it's not the data element reliability that we'd be interested in necessarily, it would be the data elements validity. So the claims data so can we feel assured that what is listed in the claims they are accurate. That's the data element validity question.

(Bijan): That's they are to me, they are.

(Paul Comaske): Yes, this is (Paul). I was actually they, I thought they did a pretty good job. They relied on the literature for validity of using claims data documents. The diagnosis of stroke and then they performed some auditing HR review to confirm the validity of the claim space approach, to validate the diagnosis of dizziness. So I thought that from that point of view, they did a good job.

(Sherrie): So (Karen), can you clarify whether we're on score level or what data elements do you call it reliability because I'm confused.

(Karen): Yes. It's tricky. So I'm sorry Michael, I'm stepping in here.

Michael Abrams: It's okay.

(Karen): For score level, they did their point estimates with confidence intervals across three years. I think.

Michael Abrams: That's right so...

(Karen): So that's what they have described as score level reliability. So one of the questions is, would you accept that, and then kind of the corollary is if you don't accept that then I'm reminding you that if you do accept their data element validation then that could suffice. Is that make sense?

(Sherrie): Okay. I'm back to concern about the amount of variability and given they only have four hospitals. There's not a whole lot that you can do with that limited sample size of the facility level. But there is a fair amount of variability over time within hospital.

And so my concern is that for example, hospital B has a visit rate of 27 that goes all the way down to two over two time intervals. And then other hospitals go from negative to positive for any level about 10 and, you know, that's a fair amount.

There's not a lot you can do with this but it's disturbing enough to be concerned about the within hospital variation. And that's why didn't give it a very high. I gave it a low rating on that score.

It would be helpful to know, you know, obviously there's only so much you can do with the sort of, this almost looks like pilot data with the data you have on the other hand. Boy, that's a lot of variation and it's a little disturbing.

Michael Abrams: Yes, other comments about that reliability testing or whether or not it indicates stability. So the last person spoke, express some concern about that.

(Bijan): Yes. I think, this is again this is (Bijan) and I had the same comment. And I think for Hospital B, they did mention that there were some transition during the major periods, but even if you ignore Hospital B, then Hospital C it's kind of goes from 19 to 37, almost it doubles.

So, although the developer is sort of saying that, you know, there's not enough, it's not, you know, the numbers are not heterogeneous. But I think, you know, when the numbers become, you know, double, I mean you kind of have to wonder whether it is really stable enough.

Man: Hi, I also was very concerned because even the center was upgraded with stroke care, and it went from 27.5 down to 2.3. The 2.3 still fell within the confidence intervals from the previous period. So, you know, how are you going to distinguish performance?

Michael Abrams: Yes.

Man: I think dramatic it was ostensibly dramatic change which still falls within the confidence intervals of the previous metrics, you know, its previous testing, sorry. So that that variability can concern me.

Michael Abrams: Okay. And so, go ahead.

(Jeff): Sorry, this is (Jeff). Can you just clarify again the rating process? So let's say that we have consensus that the measure score reliability testing either than the methods weren't appropriate or the results were not compelling. So you're saying if that's for finding them that's okay because we can rely upon that validity testing of the patient level data, which I agree was very well done. So do we not score reliability at all in that circumstance?

(Karen): No, in that circumstance you would just use, what you would score for validity in place. So it's kind of two different questions. If you think the methodology itself wasn't appropriate, then disregard. But if you're okay with the methodology and the results are really causing you some pain then you may not want to use this validity results instead.

Man: I think it's hard to tell on reliability right with four hospitals and a bunch of very, I mean, I guess you could say, you don't think this is reliable at all. You know but that's another issue I have the score, you know, it's like a two dimensional analysis that somehow summarized in a one dimensional number.

Okay. You got, you know, one dimension of the methods and then another dimension is that result. And somehow we have to like collapse that into one number but I'm still so I think what I would like to do is be able to score this

low on reliability, which is what I did. And then, you know, it's modern validity, because day element validity is pretty good.

Now what you do with that, you know, you can decide. You can decide to ignore my reliability scoring because you're just going to take the data element validity results. And that's going to be good enough.

(Karen): So what will happen is we will ask all of you to revoke on both reliability and validity. And we will take the majority vote. And the majority vote whether it's pass or fail. If the majority of you believe that the measure does not pass reliability, then we will not at this point forward it on to the Standing Committee.

We will send it back to the developer describe, you know, the concerns that you had and hopefully the developer in that case would maybe do some additional analysis and bring it back next cycle. If you guys do have it as a group then we will forward it on to the Standing Committee along with all your suggestion and they can do with it as they want.

They may just take it as it is and maybe add some additional commentary. They may disagree with your votes and add some additional content, he knows what they would do.

Man: So what you're saying is because I view the data element validity is moderate, I should change my reliability score to moderate.

(Karen): I am not necessarily saying that. I'm saying that if you think the methodology for reliability is appropriate, then you need to consider the results and decide whether those tell you something different than what you wouldn't land on if you looked at validity.

If you think the methodology for liability is not correct, then it's kind of irrelevant what the results are, right? So you would just kind of disregard it and pretend they haven't put anything down. In which case, our criteria say that it's okay not to do anything else, you can just rely on the validity piece.

(David Newman-Toker): This is (David Newman-Toker) may I just offer a quick aside, just to give people some context?

(Karen): I would like to (unintelligible) (David), just to maybe help us get out as a whole.

(David Newman-Toker): Well, so this's obviously a difficult thing because people haven't submitted this kind of measure before. And so we were trying to work with NQF staff this is to figure out what to do. We realized that the sample of hospitals that we have and the variation that we have, both in terms of the type of hospitals and the number of hospitals and everything else is not ideal. We get that.

So, but we were encouraged to submit something for level reliability and so we did. We knew that technically, we didn't have to sort of meet the minimum standard. We agree with you we concur completely the validity approach is the place where this is strong in terms of the data element validity. And I hope that you'll consider the sort of the context.

We actually had one other set of data from the Kaiser Mid Atlantic data but technically it was a health plan and not an individual hospital. So we saw, we have more data than we have, but we couldn't really submit them because the plan data then was only one plan and we didn't have multiple examples of a plan.

So I think what we were trying to do was follow the recommendations in terms of what our submission look like. But your assessments are very much in line with our opinion of the measures so.

(Sherrie): (Karen), this is (Sherrie). I'm now mightily confused because it looks like to me that the reviewers, the developers did, there is a correct analysis, they did it in the right way. But they're being penalized because there's not, there's a fair amount of within hospital variation. If this is being used to assess variation between hospitals that's a problem.

Now and then leaning on, oh yes, but the data sources is good, is to me a real logical inconsistency. It's like okay, yes, but we got it from the right data source. Okay. But if you can't tell if it's being used to compare hospitals, that's focused, that's what we're where you should see more between them within hospital variation to do it reliably.

So I'm a little bit, you know, I'm jarred by, we're talking about penalizing the developer for doing the right thing, but finding some really concerning findings at least initial ones because it doesn't mean look like we've got enough data for yet to make some more long or more sweeping generalizations about how reliable this measure will be in the field.

But I'm confused about what the action to take is because they do report the scores. They're not, they are concerning. They did it the right way and but now we're going to ignore that and say oh yes but the data source is good. I'm really that feels me like a jarred consistency.

(Sherrie): So don't ignore it if you think they did it the right way.

Michael Abrams: So here and this is Michael speaking. So here's one thought I have about the way they proffered their reliability. There's a time component in there that is leading to variation. I actually think they would have been better off just doing a straight split half and they would have come up with better indication of stable results.

So, perhaps that was the fatal flaw to the way that they proper that reliability. But there certainly is variation at least in that. I think with hospital B showing that with time never changes, appearance. So they couldn't, you know, then say, you know, things stayed stable in the same hospital over time, so.

(Karen): So, that's is not exactly. There's variation across time within hospital but there's also variation within hospital within a time period. And that's what you're looking at it that's how it's going to be used. This is going to be used at a point and time to compare hospitals, the confidence bands there too wide to see that you could use it for that purpose. So, this is back to how are you going to really use this in practice.

So across time, there's a lot of variation and within hospital, there's a lot of variation, so both are matter of concern.

(David Newman-Toker): Do we want to look at validity maybe quickly and then vote and then revote or what, sure.

(Karen): Well, we're going to be doing voting offline with this service. We don't have to take any time. If you have that open and you feel like you're ready to vote on a liability so go ahead and do that now. So you don't forget, otherwise to vote your idea so you don't forget later on when you do go back.

(David Newman-Toker): Yes. So, let's move on to validity. Let me so it's the bottom of the first page ratings of validity. Let me just quickly review for you how that was done. It was data element testing with both ICD-9 and ICD-10 codes implicated, which is good. And they used a previous work to look at sensitivity and specificity regarding the numerator which is stroke events in particular.

The results they're listed for you on the beginning of the next page. They also look more recently using that analytic reviews to validate various aspects of the testing as well. And again, reporting sensitivity, specificity, how do you predict the values that were relatively high, although notably low for one particular code or set of codes referring to stroke.

And then they did specific chart reviews and I'm getting into the middle of the next page there with regard to or electronic health record reviews with regard to identifying dizzy cases, again, both ICD-9 and ICD-10.

The numbers are presented there. Perhaps this might assuage the concern about venue perhaps it doesn't but they did do at least within their Johns Hopkins facilities demonstrate very high correspondence with regard to discharge status and identifying particular cases of relevance to this measure.

And then talking about the validity of the measure to discriminate between hospitals, they also demonstrate that looking at across the four facilities. The rates were and this is a second to the last bullet before the items to be discussed. You can see across before facilities, the rates of dizzy visits vary from 2.7 to 21.5 per 10,000 dizzy visits.

So that's they're suggesting that there is indeed variation across hospitals in this case as opposed to demonstrate instability. And then finally, the last

bullet there before the items to be discussed point is that they talked about this out-of-network issue and that there are something like 17.3%, it's a little bit unclear what they mean.

Presumably they mean 17.3% of their total cases are lost or part of the data is lost related to the fact that it's an out-of-network hospital where they can't see the data even with crisp using the Chris system as well because of the way it's coded in the Chris system with regards the primary diagnoses versus secondary diagnosis.

So that potentially is a threat to validity in the sense that, you know, you're missing a bunch of cases where you want to see the follow-up and the follow up my change your result. So I'll pause there. And, you know, key point about the validity perhaps is that missing data piece but open it for a discourse among you all regarding your concerns or questions about validity.

(Paul Comaske): So this is (Paul). The said item element validity seems strong, the measure of validity was a little less clear.

(Sherrie): This is Sherrie, I agree with (Paul). I think the issue of variation within hospitals again because when you're comparing mean rates, the average rate per 10,000 visits. The confidence intervals are very wide in these comparisons and when you're trying to compare one rate from another that's a problem. And it's going to be a problem looks like for this measure throughout because you can't get significant differences between hospitals with confidence intervals that wise.

(David Newman-Toker): Sorry, this is (David) again. We did see a statistically significant difference between our hospital that orders MRI all the time and our community hospital that doesn't have access to any stroke services. I realize

it's not as good as having, you know, 500 hospitals and really seeing the variation. These are all you know, Johns Hopkins Hospital and there's going to be less variation necessarily. Hopefully that at least helps.

(Sherrie): Yes, but the effect size is so huge that, you know, it almost it's strange imagination you can get in the fax that big. I mean I agreed that when you go out and take this beyond four hospitals, you probably are going to get some outliers that really you could identify some performers.

But the confidence interval for still disturbing. When you're just, well, maybe this isn't the right place to ask. I was going to ask about adjustment. Did you do anything trying to narrow that confidence interval?

(Karen): You can go ahead and answer that (David), this were in validity.

(David Newman-Toker): So, you know, this issue of adjustment, I know you guys had a question before, but whether there was adjustment or not. We didn't know how to describe the observed minus expected methodology. Originally, I said it was adjusted. And then we had a back-and-forth discussion with the folks that you're working with.

People generally felt that it wasn't adjustment in the sense that you have thought of it that way. So there's a little bit of terminological confusion there, but we do believe that we account for the base rate differences in the populations because we're looking at patients have their own controls or population of patients does have some control.

The issue of whether we took specific, make specific attempts to try to narrow the confidence intervals, most of the things that we tried to do in terms of parsing out the liability and everything else naturally sort of broke it into

smaller pieces that happens in your host weren't any better. We didn't do any specific methodological things, trying to narrow the confidence interval in an adjustment process was there a particular method that you had in mind?

(Karen): Currently, not specifically but, you know, did you did try - did you do any hierarchical modeling just, I know the sample size is tiny but just to see if you could kind of get the hospital level effect and take it from the (unintelligible).

(David Newman-Toker): So good, we did not, again, sort of sample size constraints. But I think that what, you know, what we have here, you guys don't know this should have behind the scenes but what we've been doing for the past two years, contextually. But we're trying to take a more complicated approach where we plot out the rate of return curves for these patients.

And we have much more information conceptually over time and condense it down into a measure that other people can, you know, use by counting relatively speaking rather than doing something more analytically intensive, even though I realized that the sampling issue makes this perfect in terms of it difficulty level. It's much easier than the stuff that we're trying to do, that we're doing for our own purposes internally.

We were trying to make it simple. I realized that in the process of doing that, it's imperfect in terms of the noise issues. But the other problem here is that if we can't, if we can't get this out into the public consciousness, we can't get more data it's sort of a catch-22 problem for this sort of new metric issue.

Michael Abrams: So, Michael speaking again here, so any other discourse on validity and anybody want to make a comment about that the missing data, the out-of-network missing data issue? Okay. We'll move on to the next.

(Karen): Is there anything else?

Michael Abrams: I don't think so, no.

(Karen): Does anybody else has any other. And let's be clear, when we say missing data is the idea there is and (David), I'm asking you because I want to make sure I understand this measure. The idea is that a patient can go to any hospital when they have a stroke, not necessarily just the one that they had that dizzy diagnosis earlier.

So that missing data businesses saying hey they went through a different hospital when they actually have their stroke. So it might be hard to know if there was a stroke after that dizzy diagnosis, is that correct?

(David Newman-Toker): Yes, they're two - yes, absolutely. There are two layers though. The first layer is not knowing whether the patient, you know, who was seen at Hopkins Emergency Department and discharge got admitted with a stroke, the University of Maryland.

We address that broadly by dealing with Chris regionalized data and the Kaiser plan, which you don't see the data because we first not submitted was, you know, they have their own follow-up for all their patients. So they have sort of all the data.

But the second piece of that is that the reason why the Chris data are incomplete is that they only rely on, they rely on Maryland's hospital rate setting condition, which only gathers data on hospital facility-coated diagnosis, which turned out to be more different than anyone would like than the professional fee coded diagnoses and it's the essentially they're complimentary data set.

So when you use the professional P code of diagnosis and the facility P coded diagnoses, you get a higher event rates and using either of them alone. And that the Chris data don't have those out-of-network professional P code. They only have the out-of-network hospital facility codes and that's the 7% key percent that are missing the professional P Codes from a Chris data.

Michael): So just be clear, (David), if you would, there's 17.3% of your total sample that you're losing or of the out-of-network sample? Can you clarify that for us?

(David Newman-Toker): I'm trying to find the spot. I know we have (Matt) on the phone. (Matt), do you have that in front of you. I'm not finding it.

(Matt): Yes, I need to look at (unintelligible) too.

Michael Abrams: Okay. So that remains is a question. You know, that one in five of all the observations that you're looking at, or is that one in five of a subset of the out-of-network that would make a difference obviously and some idea about the impact on the measure as well, to the extent that you can provide that?

One other point, perhaps before we close out the discussion. One of my colleagues will line up, any further interest among the committee right now to talk about the risk adjustment approach that they use here, which, you know, again, was looking at the 3 to 12-month window post ED discharge to look for baseline events? Anybody else wants to discuss that whether that was appropriate or of concern moving forward here.

(David Newman-Toker): So, can we move on to the next closure, the inspiration here?

(Karen): Okay. I think we've discussed as much as we can. (David) is or (Matt) is somewhere on the next hour or so, you find that number and can answer that…

(Matt): I have it here. So the estimated missing stroke events divided by the total stroke event so it is 17% of all numerator event.

(Sherrie): This is Sherrie. I just, I don't want to bog us down. But if you get a sense of is missing it random or is it missing not at random that would, it's probably related to the out-of-network issue.

But your assertion that it reduces precision to do some kind of multiple amputation, it actually does not, it does reverse. So it inflates the meaning narrow the variants. That's why I was asking you about sort of sensitivity to treatment of these kinds of, you know, adjustments.

And thought this is not a risk adjustment but if you did multiple entities, did you think that doing multiple amputation and looking at the sensitivity of your results to that that sensitivity, specificity of your results to that kind of amputation?

(Matt): We did not do that. So I'm sure that we could, we could. Again, I think it's a good point, we should assess whether the - missing this is likely to have an impact.

I think the conceptually from our perspective, the relationship between the out-of-network follow-up missing is, which I think is critically important was more significant than the relationship between the hospital versus professional fee billing.

So conceptually, just from a structural standpoint, it seems to me that like (unintelligible) it was a reasonable assumption to make will say that the professional fee and hospital fee billing relationship would likely be similar overall across hospitals rather than different. There's no way for us to verify that 100% certainty with the data that we have.

But conceptually it seems much more likely that there's a bias in who returns to another hospital but not necessarily a bias in whether a code or codes the facility code as stroke or not as a poster physician because of the stroke in that.

(Karen): Okay.

(Matt): Sorry I didn't answer that question well. Did somebody said, they were still confused?

(David Newman-Toker): Well, that's okay. We will clarify in the future.

(Matt): Okay.

(David Newman-Toker): Thank you.

(Karen): Okay. Thank you (David). So hearing no other, I'll give you just a second. Any other things you want to ask about this measure? Okay. If you have your SurveyMonkey open, you may want to go ahead and vote on validity as well. We'll be voting on both reliability and validity for this measure.

All right. Let's go to the next measure. So that went …

Man: So excuse me, I just have a question on the SurveyMonkey issue. Is there a concept that as of the discussion where we will have reached consensus, and therefore we're all go the same or you're just going to reach our, you're going to tally the votes? And then in other words, what happens if we don't have consensus on the SurveyMonkey after all of this discussion?

(Karen): Right. If you don't have consensus, even after all of this discussion, we will go ahead and forward the measure to the Standing Committee. And let them grapple with all of the description of what you're struggling as well. So that will go to them.

Man: They will know what it was that was of concern in this committee?

(Karen): Yes, absolutely.

Man: Okay.

(Karen): Yes. They'll actually know it in two ways. Number 1, we will write a fairly detailed summary that will go in what we call their preliminary analysis that we prepare for them. But we also provide your initial analysis at PA that you guys did. We make that available too, so they could actually go back and look at, you know, your actual words, not just their summary if they want to.

(John Bernot): This is John Bernot. I don't know if others ran across this but I voted on this measure and I couldn't get back to, you know, it seemed like I had to leave SurveyMonkey in order to get back into vote on the second measure, is there…

Man: That's exactly what just happened to me.

Woman: Me too.

Karen: Yes, that was done by design. So you'll have to spend a separate survey for every measure you vote on today.

Man: All righty.

Karen: Did you guys hear, did you hear (Miranda)?

(Miranda): Yes.

Karen: So that's how it'll work. You have to vote and then get out and come back. Okay, let's move away from dizziness and stroke and vote again, so.

(David Newman-Toker): Thank you very much for the opportunity to be in and listen and to comment. Appreciate your time. Thank you.

Man: Yes, thank you Dave.

Man: That was helpful.

(David Newman-Toker): Bye, bye.

Man: Could you please say what measure number we're moving on to, there's a disagreement between the agenda and the discussion guys and to the sequence of these measures.

(Karen): Yes, we are doing 2549 E gout serum your right target. Have you found that in the guide, everybody did?

Man: Yes, thanks. I'm looking right now.

(Karen): Okay. So this is a new measure but it's not a completely new measure with NQF. It is an easy QM or an e-measure that has been previously approved for trial use. You guys might not be as familiar with that kind of pathway to considering measures. So, I've added a little bit of detail about what that means here.

That does not mean that this measure has been previously endorsed. It has not been approved for trial you. And that means that some of our other criteria, a standing committee has looked at the measure and like it met the other criteria. But by definition, testing data reliability and validity was not considered previously at all.

So this is the first time that people will be looking at reliability and validity for this measure. So it is the percentage of patients 18 and older with a diagnosis of gout treated with a new LP for at least 12 months. And then the numerator is most recent serum your rate result is less than 6.8 (unintelligible).

It is an intermediate clinical outcome. The data sources are EHR and registry. So again remember this is an E measure or an ECQM. And in terms of it is not risk-adjusted, level of analysis is noted as those group or practice and individual clinician so both has been selected.

In terms of reliability, the ratings of the sub group were four high and one moderate, which means that it does ask with a high rating. We're saying high because that's the majority rating.

And they did signal-to-noise analysis from the rise registry data. The results are there in front of you. We did note that because this measure has been specified for both the individual clinician as well as the great practice that we should think two sets of reliability results, and one for individual, one for group.

So we did not see that we are going to ask the Standing Committee about that and what they feel is appropriate there. The developers also described some testing and they actually put it in their submission form is reliability testing. However, what they did, we actually consider it to be validity testing of the data elements.

So I'm going to go through that even though we're not going to discuss reliability necessarily just because I think some of what they did hinge on their validity testing. So they did two different things.

They looked at three sites and actually use the ECQ specifications to get an automated report. And they checked back from what's on that report, check back to the actual medical record in the EHR and computers and Kappa agreement scores. And they did it for the numerator the denominator and the exclusion. And the results, the capital results range from reflect the lowest of the three of those as point seven, eight all the way up to one.

They also did, they talked a little bit about some auditing of the rise registry. So rise registry actually takes data from many different EHRs and pulls the data in and they did an overall count of correct and incorrect values, they looked at 644 patients across 13 providers. So this doesn't exactly conform our requirements for data element validation. And so they did part of it, they really needed to not just give you one overall number, they needed to and split it out between minimum numerator, denominator and exclusion.

And we also asked for something more than just percent agreement so some kind of Kappa or preferably sensitivity, specificity, et cetera. That doesn't mean you can't at least see what they found when they looked at that data. It just wouldn't stand on its own for our testing.

Now, when it comes to validity, we actually had consensus not reached amongst the Subgroup. So, two voted moderate can be low. And they did the data element validation which I have just described. They also did some score level validation. And they talked about a face validity assessment. And I think that face validity assessment is one place where people really got a little confused.

So in terms of their concern in validation, they correlated their own measure results with another measure that can be computed out of rise registry. It's a monitoring measure.

And they didn't really talk about why they chose that as a competitor or the results they expected to see. But they did have a high correlation. And once again they should have split that out between individual versus group. So that's the caveat there.

We have pre - even though face validity is -- we consider it kind of lowest form of validation first of all. But we do have some kind of requirements for it. They, I think talked a lot about what they did and who looked at it. But they had, they did some face validity from two different the out measures. And it doesn't seem to be the one that is actually under consideration, not quite sure who the experts were, what the scoring range was.

And in that whole assessment, it was really a little bit more about or at least part of it was about the development of the measure and a lot of different disagreement it sounded like as to the threshold uses the measure of the 6.8 versus the 6.

Now with all that in face validity, (unintelligible) to remain you that face validity is kind of the lowest form of validation. So what we really - we don't want you to necessarily to ignore it completely. But really pay attention to the empirical testing that was done. So the correlation analysis that you had as well as the data element testing that you have.

And don't get too in the weeds on that face validity because I'm not quite sure how much that matters in the face of actual empirical results. This measure is not risk adjusted. They did provide a conceptual rationale as to why they did this.

So I think I've gone through most of all I wanted to remind you all that we're less concern with face validity if empirical testing was conducted. One thing that did come up that was a note from one of the Subgroup members, so it was about the competitor that we used in the construct validation. And right now in terms of our requirements, we are silent on what measures are appropriately used the score level validation analysis.

So, we do not necessarily require that they're using an outcome measure, for example. So we're silent on that. And basically the idea is, is there enough rationale or do you feel like what they did use as a reasonable measure to this for their validation.

So in terms of the action items that we want to have you discuss, we are not going to discuss reliability unless any of you feel like we need to. Because

again, everybody pretty much landed on the side is reliability, reliability was okay.

Any, desire to discuss reliability?

(Paul Comaske): There was one issue regarding the specification, which was a little bit - they just need to clean it up. It says there is several points include the algorithm for how the data is collected, there's an exclusion for solid organ transplant.

However it's certain of the text, there's also an exclusion for Stage 3 to 5 renal failure. But it's not specified throughout the application. So it's unclear whether that is part of it, the exclusion or not. They just need to, you know, clarify it.

(Karen): Okay. Thank you (Paul). So it was kind of one way in one park and another way somewhere else in the submission so unclear.

(Paul Comaske): Correct.

(Karen): Thank you.

(Sherrie): Karen, this is Sherrie. I voted moderate for validity. But I have a question because I didn't know what to do with it.

You can't get a result from the gout serum, your rate monitoring without monitoring. Those two measures are technically not independent. They're pretty much in the same thing. So how do you guys deal with that? For example, performing hemoglobin A1C and then you get a value.

Well, you can't get the value unless you perform hemoglobin A1C. So how do you guys deal with the issue of validation when the measures aren't independent?

(Karen): That's a good question. We don't require necessarily that they'd be independent. But I think part of your, you're not going to like me that they share it. Part of your decision today is whether you think the dependence is so bad that it's not even worth kind of thinking about the analysis. You know what I'm saying?

(Sherrie): That's what I'm asking because you can't put a value unless you perform the test. So, you know, those things are pretty much the same almost. And if you're measuring - if you're using one is evidence of the ballot, the accuracy of this measure for comparing folks being compared. And those measures are not independent and I'm not sure what to do with it.

(Karen): So, in that case, you may want to just completely disregard that score level validity. Again, this is just a plain, intermediate clinical outcome measure. So we don't require that they do score level validity. So you might say, hey, they did it that, you know, really, really weak, and I'm just going to ignore that.

(Sherrie): But then you're back to face validity, which you said was not optimal in this case?

(Karen): You know, actually, what you're back to is what they described under reliability. So what they talked about under reliability testing, they actually put in the wrong spot. They should have put that under validity. So they looked at the three EHR sites, which doesn't sound like a lot.

But I will remind you this is an E measure, and we - it's one of the few places that we actually have requirements in terms of how much testing has to be done. And our requirements are E measures have to be tested in more than one site. So they've actually tested in three sites. So they have hit that one.

So if you go up to the (unintelligible) number these pages, many of the number of these pages so we have page numbers on these discussion guide. You can look at the Kappa scores that they computed when they compared the EHR automated report to the full EHR. So, you can look at that whether now or in addition to if you'd like face validity.

(Sherrie): So, you're considering that as criterion the physician charges of section is criterion validity?

(Karen): Yes, for the data on it.

(Sherrie): Action is the criterion to which the MR obstructions being compared. In that case I couldn't find whether how many physicians that have extracted the charts and how much agreement if more than one was, there was between the physician?

(Karen): The physician who abstracted with chart, I'm sorry I don't, I'm not following your question, Sherrie.

(Sherrie): How many physicians abstracted charts 1, 2, 6 and if more than one, first of all, if it was one, how accurate is that individual? And then secondly, how If more than one what was the agreement between physicians abstracting the chart?

(Karen): Do we have either of the developers from this measure. This might be a good time to directly ask the developer, if anybody is on the line.

(Tracy): Hi. Yes, you have (Tracy Hansen) and with (Lisa Tudor) from EACR on the line.

(Karen): Hi, Tracy. Can you can you give us a flavor, do you understand Sherrie's question and can you answer it?

(Tracy Hansen): I just want to make sure I do understand it correctly. You are simply saying for the abstractions that we did at the three testing sites for the measure. You want to know how many physicians were involved in that?

(Sherrie): If you're comparing the MR obstructions to the physician, chart obstruction. I just want to know how many physicians abstract were involved in that chart of section 1, 2 and.

(Tracy Hansen): Let me look through our documents, I know that was three sites but I don't believe it was, I believe it was more patient-focused and then specific provider focus. I'm going to have to double check that real quick so.

(Karen): So, we'll circle back to that one. I think Tracy is looking for the number of providers that were included in the extraction of data. But you're, I think Sherrie, you're asking whether or not there was more than one individual doing the medical record abstraction that.

There is only one physician per site to my understanding that was actually checking, pulling the medical records and reviewing the data per site. And Tracy can verify that.

(Sherrie): Okay. So the physician in the side confounded then. All right. So, each physician is independent, but you didn't give them a course set of things to abstract and see how much agreement there was between physicians about specific kinds of issues that could come up in the chart extraction process itself. You didn't do that, right?

(Karen): So, I'm not sure I understand your question. But my understanding is for uniformity there was a single abstract or at each site.

(Sherrie): Yes, my concern is when you don't know how accurate somebody is, if you're using them as the criterion that sort of needs to get established. You know, how good are then, we assume that all physicians are good and are perfect at what they do. But they're probably not. And there would be some disagreement between physicians if you give them a common set of things to abstract from a chart across the three sites.

And what you want to do interpret that as criterion validity is know how much agreement there would be across the three site physician so that you could tell, you know, what you're looking at in terms of the variability of any about what they're drawing out of the chart. Are they agreeing in general so that you can use the physician report as the quotes criterion? -- I get it.

Woman: I think it's helpful to have that feedback. I also think in terms of this is verifying that a lab result is a lab result. So I think we had some assumption that the physicians would be capable of doing that consistently. But it's helpful to have your input on, on doing more formal criterion validity setup.

(Karen): And this is Karen from NQF. Just to be clear, you didn't give these physicians any specific training or anything along those lines, you just ask them to compare?

Woman: We asked them to verify that each patient that was identified as having a serum urea level within the measurement timeframe below, at or below a 6.8 milligrams per deciliter actually have that value documented in the medical record.

(Karen): Okay.

(Tracy Hansen): So, this is Tracy I do want to point out here too that in that they were given, you know, the data collection forum with very explicit questions for these elements and saying yes or no and making sure that that match with the HR with what was extracted from the record as well.

(Sherrie): The reason that I asked is we did a study and actually found that the physicians don't agree with MR. The physicians actually, there is a fair amount of mismatch in prostate cancer. So I was just curious to know if you guys have actually gone through the process of agreement between physicians.

(Karen): Thank you Sherrie. So we talked about the data element testing it. Anybody have any other questions or concerns about the three sites, EHR testing?

Okay, the score level testing results. The question is just whether, you know, you feel like that those are compelling or not. It sounds like Sherrie thinks not since they are very much dependent on each other. And then, the face validity assessment again, we don't even allow that by itself for EHRs.

And, you know, the quibbling that happens, I think in the face validity assessment about the 6.8 versus the 6 is more clinical. So we would want the Standing Committee to deal with that anyway if there's anything to kind of

fuss about on that one. Would anybody like to talk about the face validity assessment?

Okay. There's a little bit of concern about not having exclusions testing of exclusions done. And so our question for you is did you see enough information provided around exclusions to decide whether or not there are physician frequency towards inclusion or specifications on the measure?

Perhaps the other way to ask it might be, is anybody concerned with the exclusions and what you were told about those. And perhaps the other way to ask….

Man: Can you tell us what question on our PA forum is in regard to exclusions?

(Karen): Let see. I might, came in a minute on the testing attachment it is Item 2B.2.

Man: It's under the form I'm struggling to find out what a whole lot, what I said about exclusions because I can't find the question on our form that we complete it right now.

(Karen): Yes, give me just a second. I'm still looking for it. I haven't quite found it. The word exclusion says it quite a bit. It's a question as well. Please describe it (unintelligible) with measure exclusion.

(Jeff): Thanks. I just noted, I had no concerns so…

(Karen): Okay. I think sometimes people aren't quite sure what we're asking for and they're awesome what developers provide to us is a frequency of exclusion without, you know, major testing. And I think sometimes people are looking for maybe sensitivity analysis or something along those lines.

We don't often get that. And let's see, I can tell you, let see, (Jeff) I think you might have been the one who was mostly concerned about exclusion. And then (Paul), going back to your specifications question.

Man: I think my only concern was there wasn't testing.

(Karen): Okay. So, since we're not really and we don't give any instructions on what we might mean by testing. They did tell us what the frequency is exclusions so there are number of patients who were excluded were from the three sites. Is there anything you'd like to the scene for this trial…

Man: Does that matter? I mean, the reason, you know, I mean just the number of exclusions doesn't answer the question of whether the exclusions are somehow distorting. But like I said these are not really any guidance about that, you know.

(Karen): Yes. Okay.

Man: I was never, a lot of the registries kind of make a statement along lines of thing with regard to missing data. You know, there's not any missing data because we don't accept it. And I always find that a little strange. I mean just doesn't mean is missing data is that you don't, you just delete it.

(Karen): Okay. How about any concerns about the lack of risk adjustment? I think, mostly relied on a conceptual rationale for not risk adjusting. They gave a little bit of information about geography. Is the conceptual rationale sufficient or did anybody want to kind of talk about lack of risk adjustment?

(John): This is (John Vibes). The comment I made in across the 10 measures. I probably made it in six or seven of the ten. Just while they presented a rational I disagree with their rationale. And I didn't have copious notes typed in for a variety of reasons. But I think my comment here was it seemed like and then this is going off of memory which is dangerous that it appears there's geographic variation.

And unless we control for that apparent bias providers in a certain geographic area are going to tend to look worse simply because we're punishing them for being in that geography. So it seems like that they need to figure out how to control for that bias so as not to punish a facility who happens to be in certain part of America.

(Karen): Tracy or Lisa, do either of you have anything you'd like to tell John along those lines in terms of geography and why you didn't feel that you needed to response...

(Lisa): This is Lisa. And I'm going to ask Tracy to jump in because I am not as familiar with the geographic testing. And I don't actually see geographic testing in the testing form that suggests that there are geographic differences in performance. What I see is geographic distribution of practices. So those percentages in Table 7 are not performance. Oh, I see bottom up to 25%, top 25%.

So there are some geographic differences. I personally don't think geographic differences are used for risk adjustment. I think the argue for argument that I have heard in contrast to our conceptual model for risk adjustment is that some practitioners care for a disproportionate group of patients that might be challenging to have their serum urate levels appropriately managed.

It's our experience that the combination of the particular serum urate target that we hit, we identified as part of the outcome definition for this measure, along with the abundance of medications including newer medications in this area suggests that there is much less of an argument for risk adjustment in this measure than there have been previously. So, there are more medications and available to more types of patients than previously.

And we are, I think we have chosen to identify a serum urate target which is much less likely to be as challenging to hit than a more stringent targets. Based on your underlying comorbidities, or other medical conditions, or even social risk factors in terms of access to medications and things like that.

So it was the panel workgroup for the measures determination that with this combination of factors, it was better to move forward with a measure that Lisa think is critical for improving gout control.

Then to create a risk adjusted measure that might be more challenging to implement and may not provide us as rapid information as this measure would in terms of under treatment of symptomatic Hyperuricemia in the United States.

(Karen): Okay. Thank you for that. I think the only other thing that came up is the Subgroup did notice practice level variation. And the question is speaking of that there is two differences between providers and just to note that we're really talking about quantitative differences, not necessarily clinically meaningful differences. So that can be discussed by the Standing Committee.

Woman: Karen, can I ask a question? This is this applicable of the last issue sort of. But I'm sort of a recommendation maybe but a thought that if you did the proportion facilities or the proportion of groups that or individuals that did the

serum urate monitoring. And had an in-target value versus those who did the serum urate monitoring and did not or at least the proportion that had we're not in target.

It gets you out of this, you know, the independence issue I was worried about. And it looks like it's kind of resonating with the intent of the measure the way it was just articulated by the developer. So if you did the test and if target value get credit, if you did the testing, it was not in targets then you don't get credit. So it's kind of a way out of that independence problem. And it does speak to, you know, how many people were above and below the target given the test.

(Lisa): That's helpful. This is Lisa, can I just ask a question in the intended denominator for that combined measure? Just so I understand your suggestion.

Woman: Well, could get into the, whoever the at-risk population is for group that…

(Lisa): Okay. Great.

Woman: You know, the test plus in addition, having an in-target value.

(Lisa): Great, very helpful. Thank you.

(Karen): Okay. Were there any other questions or concerns on this measure?

(Jeff): Karen this is (Jeff). Sort of a general comment, a lot of these sort of intermediate outcome measures or things that are called intermediate outcome measures, look a lot like process measures. If - I guess they're called an intermediate outcome because they're using some kind of clinical lab value.

Woman: Yes.

(Jeff): But the clinical lab value is completely determined by the process. So, essentially it's a process measure. Which is basically why none of them risk adjusted. So I don't think what I said matters but it is, you know, it just, I mean it's essentially, you know, they treat them like they're process measures because the quote outcome is completely determined by the process.

(Karen): Yes, I think that is an argument that's most used. In terms of our other criteria, those of you who maybe haven't sat on NQF Standing Committee recently may not realize that, well we do think maybe a little bit differently about intermediate clinical outcomes in terms of risk adjustment, whether or not, we definitely think about them the same way as process measures when it comes to (unintelligible) we are actually looking for under the evidence criterion which you guys don't have to worry about, we are asking for evidence that that lab value and what have you is important for the, you know, the health outcome of interest. And that's where we talked a lot I think about the threshold (unintelligible) measures.

Man: But the threshold just that the, all of the data and the face validity, the expert panel that references were all based on a level of six and yet the measure 6.8? I don't know why they ignored that discrepancy. I don't know, you know, I don't know urate levels well enough to know that you could safely extrapolate and get the same degree of precision.

You know, it just - everything that they presented, you know, with articles and expert opinion was a level of 6, not 6.8. And then they made the measure 6.8.

(Karen): Yes, so when the Standing Committee - if they get this measure they will be talking about that at length I'm sure. (Unintelligible) come up in validity a little bit but it mostly comes up under the evidence criterion.

Man: The face validity was all based on that.

(Karen): Right. Okay. Go ahead. If you haven't already and cast your vote for validity, for this measure again you do not have to vote on reliability.

Okay. We have almost a half hour left. So Andrew, we're going to talk about Hemodialysis. You feel like you probably make a good effort on this one in 30 minutes?

(Andrew): Yes, I think so. So we've got a set of measures here related to, mostly related to the adequacy of dialysis treatment.

And I think the concerns that were raised by our panel members are very similar or virtually the same across all the measures. So hopefully we can kind of gain some efficiencies here. I suspect we'll still have to have our second call but maybe it'll go quickly having this first discussion.

The first one we're talking about is measuring number 249 delivered dose of Hemodialysis above minimum. So this is the percentage of all patient (unintelligible) for patients (unintelligible) dialysis whose delivered dose of Hemodialysis was (unintelligible) greater than or equal to 1.2.

It was similar to some of the ones we've talked about already a lab values sort of a classified as an intermediate outcome. And sort of, you know, jump right into the concerns that were raised here that I think the concerns largely related to the reliability testing. Reliability was tested at the score level. And the

developers did assess what they call the (unintelligible) unit reliability, which they defined as the proportion of the measure variability that is attributable to between facility variants.

They used a bootstrap approach to estimate within facility variation. And one of the things that our panel members raised is that they would like to see a little bit more detail on that bootstrap methodology. The testing did result in an (entry) unit reliability score of point 808 for this measure, which suggests 81, roughly 81% of variation in the measure is attributed to between facility variation and around 20% attributed to within facility variation.

One of the, again main points of concern here is that as noted by the developer to their credit, this method for calculating entry unit reliability was developed for measures that are approximately normally distributed across facilities. And that is in fact not the case for this measure. They say it is not normally distributed. And so they suggest that the value here the entry unit reliability score should be interpreted with some caution and our (unintelligible) reviewers did, you know, (raise) some concerns about that.

So maybe we can just kind of start with some discussion about that. I think that the sort of key questions here on reliability - was the method appropriate? Especially considering that it was, the method was designed to the test measures that are normally distributed and these measures are not. And then is that - in light of that are the results, you know, interpretable, do we have confidence that that the measure can reliably distinguish between facilities?

So, any comments from our panel members?

(Paul Comaske): Yes, (Paul Comaske). I had a more basic problem with the data element reliability.

(Andrew): Yes, that was another issue that, you know, that was a concern raised that there was not any element reliability testing and there is at least (unintelligible).

(Paul Comaske): (Unintelligible) no center apart from one unit manage to complete the collection of blood specimens as recommended by the guidelines. With one exception blood collection for Hemodialysis adequacy was not performed using proper technique in any center. So this would really suggest that the data elements used for the measure may not be reliable.

(Andrew): Yes. So that's certainly something we can discuss. Was there a concern of any of our other reviewers?

(Sherrie): This is Sherrie. I don't have that specific concern. But I was also concerned about how many facilities got excluded on the basis of low volume - they didn't have patients. And for how many facilities did 11 patients constitute the entire population of patients?

For larger facilities, did you include everybody and it's not how patient samples because the issue of this, you know, those volume issues going to crop up again and again, be in over the course of these measures. So if the sample sizes all over the place by facility then we got a volume outcome, a volume issue to deal with.

(Andrew): So maybe it would - oh sorry, go ahead.

Man: And I think I was one of the ones who talked about this bootstrapping and it was about normal distribution. And they probably, you know, kind of have

been almost all for three of four majors they basically say that often it needs to be assumed or, you know.

But then I just think about I mean impacted how exactly would that happen? So that I get, you know, that need some clarification. And I think in their case, it may not be an issue because I think they still have good number here in terms of sample size.

But I guess all I needed was, you know, if someone were to do it, how exactly, you know, what exactly does it mean that when you say cost needs to be, you know, used while interpreting the results. So, that needs some clarification.

(Andrew): Great. Thank you. Yes, so maybe this would be a good time to turn to our developers. Hopefully, they are on the phone to address the questions of, for one thing, the reliability of data collection.

And then the bootstrapping methodology as well as the volume question, how many facilities were included on the basis of not having enough patients. Do we have a representative of the developer on?

(Karen): Yes. This is a subset of the developers of (UFM Tech). A number of our team members did have to leave before this discussion started. So we will have to respond to some of these issues in writing after the meeting. But I'm going to hand it over to (Doug Travel) to talk about one of the questions now.

(Doug Travel): Yes, I can speak about the relationship between normality and reliability. And we didn't discuss that, well we didn't discuss that caveat. We mentioned that caveat and but it's not the case that we're going to overestimate reliability due to the non-normality that's only going to hurt us.

In terms of contracting with the end unit versus the twin unit variability that's only going to cost us with respect to within unit variability. It's like we are going to be overestimated.

(Sherrie): Actually, this is Sherrie. It's probably not. It's actually, you know, work in the other direction because when you get tiny amounts of variability to deal with small perturbation don't matter as much. So one of the other things, I would have like to have seen is some statistical evidence of kurtosis or something that helps us understand the magnitude of the skew Number 1.

And secondly, did you think about doing like generalize estimating equations for generating the ICC because those sometimes can help you compare, if you get the same results, and at least you are more confident that your data are meaningful.

On the other hand, with this amount of positive skew. I also had some issues about how did you do the bootstrapping, how many samples with the magnitude of the samples were that you took each time. So, can you give us a little more clarification about how you approached, what the magnitude of the (unintelligible) was and how you approached - why did you use (unintelligible)?

(Andrew): So I think we did. I don't know for sure, exactly. But our sort of industry standard here is to do 100 bootstraps, which would definitely be sufficient for an exercise like this…

(Sherrie): How many patients for bootstrap? How many patients per procedure did you use?

(Andrew): So would be, yes it would be replicating the data set basically. So we want to replicate the stability of our procedure across the data set. So that would imply using a data set of the same size.

(Sherrie): No I know. So, you did a 100 hundred bootstrapping procedures of how many patients did you sample each time? You didn't - do use the entire population?

(Andrew): Yes. So we were generating data under the no using, like, what would represent the study population. Maybe I'm not understanding the question. I can think of like a quicker way to do it would have been like what's called the EMO bootstrap.

We take fractions of each facility and bootstrap that way for computational savings. Then you scale the results down in terms of variability. But that wasn't necessary for what we were doing.

(Sherrie): See that's why we wanted to kind of understand exactly how you did it because we're not, it's not clear from at least I didn't read it, what I read and what was included within facility variability. So you wouldn't do a bootstrapping procedure on facilities of having 11 patients.

But you could, you know, for a large sample of patients that say they have thousands of patients you would do a 100 bootstraps of a 100 patients each just to see how much fluctuation the results you get.

So it isn't clear, I mean, I don't think it's clear to some of the other folks who reviewed this measure exactly what you did.

(Andrew): There must be some difficulty with the way I'm describing it actually because we bootstrapped everything. There was no sampling involved. The

resampling procedure was basically regenerating data under the (unintelligible). Maybe there's a different use of the term bootstrap in this context. We're generating data under the (unintelligible) that I'm not getting across. But that's not what we did.

I also wanted to - if I can return to your comment about within versus between variability. I'm not following your logic about why we would be underestimated the within variability. That's not clear. But anyway, maybe we can move on.

Man: Any comments from our reviewers on that? Sorry, we're having a little discussion offline here amongst our staff. We're somewhat unfamiliar with that method of bootstrapping, as well. Some folks around the table here suggested that they had only know bootstrapping methods that involve sampling. I don't know if our panel members have any response to the developers' clarification.

(Sherrie): So there's one more thing, this is (unintelligible). We can follow up to this with a better explanation of exactly what we did for bootstrapping, just for clarity sake, because it seems like there is some confusion about the method - the terminology we've been using. So we will follow up with that after the meeting. Absolutely, thank you.

(Karen): And this is Karen from NQF. And I think we'd have to ask you to get back to us by COB today. We really need these (unintelligible) from our methods panelists. So if you could do that by COB today that'd be wonderful, we'll share it with them.

(Sherrie): So I don't know that we can guarantee that. I do know that there is a follow-up meeting that scheduled for next Friday the 22nd. We could definitely get

you all of the materials well ahead of that meeting so that the group can vote other, if that would be acceptable.

(Karen): So, we will get back to you with a time, because this apparently this kind of goes across all the measures. Okay, all right. So we may, I don't know if we can even ask our panelists to vote today, right?

Woman: Yes, so what we can do is look at the timeline and figure out the best date for - to give you and we'll work with you to get that information out. I think we mentioned the next meeting is March 22. So we do although at the time we want to be fair to our panel members. So they make sure of the right information.

So probably not ask anyone to vote on the measures that we discussed today. And then we'll move from there. Do that sounds fair, everyone?

Man: That's sound fine. So, you know, can I ask to the measure developer one further question actually Sherrie brought up a very good point. So I understand that the bootstrapping is simply (unintelligible) initial sample.

I guess my question then would be, so I'm looking at here you're sort of, you know, the document here you had 6,407 facilities that are at least 11 patients. So I guess my first question would be what was your unit bootstrap and, you know, did you bootstrap, like, you know, while doing the resampling did there was, you know, you need the facility or was is the patient?

(Karen): I think we'll have to check in with some of the folks that are not here right now. So we will follow up with that response.

Man: Okay, sounds good. Thank you.

(Karen): Okay. Also, could I, I just ask for the developers to give us either Kolmogorov-Smirnov or kurtosis or some statistics that looks, like, gives us a sense of the magnitude (unintelligible)?

Man: Yes, we can do that. Thanks for the suggestion.

(Andrew): Okay. With that I think kind of, again, cover the remainder of the renal measures that we're going to discuss today. So I think we sort of have to postpone that discussion until we get a bit more information from the developer. We do have another call scheduled the 18th, is that correct?

(Group): Twenty second.

(Andrew): Oh I'm sorry, 22nd. So I think we'll end up talking about this on that call that gives the developer a little bit more time to get us some information about the bootstrapping method. So I think that will probably postpone discussion of all these measures we have.

I don't think there were any concerns about validity on this measure. So we can just accept the ratings of the panel members unless anybody object to that. Anything else to cover on this side? Oh, there was one other measure there was some concern about validity.

Again, I don't know if we want to wait until the second call because we have the same reliability issues to discuss. I guess since we have some time maybe we can jump down and just address that very quickly. Measure 1423 and 2706 actually the same kind of issue.

Both of these measures actually were had only face validity assessment. And I don't know that they, I think the face validity assessments that they did probably wouldn't be considered acceptable by NQF standard.

They basically presented information on the development and approval of the measure by their clinical and technical expert panel. And sort of, you know, suggested that that implicitly demonstrate the face validity of the measure.

But I don't believe they gave in any of the, they did not do the sort of systematic assessment that we typically expect which is the rating of, you know, whether the measure, you know, accurately reflects quality and then give us some, you know, the voting results and some additional information like that.

Do we have any discussion about that for measure 1423? This is again, a dialysis adequacy measure related specifically to pediatric patients on Hemodialysis.

Man: I think can you just, this is a maintenance measure, correct?

Man: This is a maintenance yes. So, that's another issue that we typically expect empirical validity testing for a maintenance measure. And if they only have face validity at that point, we expect some justification for that, which I don't believe we saw in this mission.

(Sherrie): This is Sherrie. Can I ask a question? Because it was only like I was confused about whether it's 13 or 14. But that few eligible facilities, I mean, we can, you can get some empirical data. But what's think you have policy on, I mean, this is a very tiny number of facilities for which to draw empirical evidence.

It's pretty tough on the developer to, there is probably going to be a lot of within facility variation. So I don't know what we would garner out of getting more empirical testing.

(Joe Masana): Andrew, this is (Joe Masana) from (UM Tech). May I introduce a condiment I think that's directly relevant here.

(Andrew): Sure.

(Joe Masana): So, I'm one of the nephrologists on the clinical team at Michigan. I believe the two measures that you're discussing right now are pediatric measures. And when these measures were first presented a number of years ago there was a hot debate at the CSAC.

And I think even at the board of directors level about the difficulties of pediatric of developing and implementing pediatric measures in the East or the community because of a very small population of pediatric patients.

And there was specific discussion about those difficulties and whether or not pediatric measures should be considered in the same light because of the adverse numbers issues as adult measures. And it might be worthwhile to go back and to revisit some of that discussion about the pediatric measures. They were considered in a different context, I'll say then the adult measures that we've developed at the time.

(Paul Comaske): This is (Paul). I could sense that sort of tension because the data suggests that potentially higher threshold might be more appropriate for the kids. But then when it comes to guidelines people just decided to go with not lower than

adults. Because I guess there is just isn't enough data to be able to make a new threshold. Is that what happened?

(Andrew): That's a reasonable statement of my understanding of the work that was done back then. Yes.

(Paul Comaske): You know, there hasn't been anything to change it. But so it's still sort of in face validity phase. But, you know, that may be the best we can do or I don't know? Still connected?

(Andrew): Yes, sorry. Again, having some offline discussion here.

So I think we'll probably, what was our decision here, do we need to postpone to the next call? And this one, I think we will sort of started another discussion here so we can continue that on the next call.

Was there anything that the developer might be able to clarify in the same way as they are with the reliability testing or maybe what we might ask to the developer on the NQF side is if you could give us some justification in writing for submitting this measure with only face validity testing for maintenance review.

Because that something that we typically expect that they give a rationale for why you or not able to do empirical validity testing at this stage. So it might be useful for us to have some, you know, just a brief justification for why that you're submitting this with only face validity testing.

(Sherrie): Yes, this is the developer we can do that.

(Andrew): Okay, great.

(John Batis): This is (John Batis). Just two comments on this sort of topic. You know, it seems like, you know, we're being told in the guidelines - in the guidance that, you know, after maintenance, this measure must go beyond face validity. So, right now, our hands are tied. We have to rate it as accordingly and that would be, you know, failing.

Secondly, in regard to almost most of the instances we've seen in this round of these 10 measures, if the measure developers said we did face validity, they essentially provide virtually no explanation of what process was involved in that face validity - just said they used a clinical panel and did face validity. That doesn't give us a lot to go on to give us any kind of confidence that the, you know, the integrity in which the measure was vetted through face validity.

I'd like to see more stated about face validity other than just, you know, we used face validity testing I think, you know, there's some acceptable face validity processes. And there'd be some processes that aren't acceptable. I think it needs to be fleshed out further.

(Andrew): So if there are any clarifications that the developer could give on the process for assessing (unintelligible) as well, that would be helpful. And with that I think, sorry go ahead.

(Sherrie): It's okay to say we can, we'll provide that and to that in the write up.

(Andrew): Okay. Great. All right, well that I think we can probably close out our call. Do we have any next steps or anything else we want to talk about with the panel?

(Karen): No, we'll send out more information about when we get the information and what they can expect on the March 22 call after the fact.

(Andrew): All right. Well, thanks everybody for your time. We appreciate it. And we will talk to you again soon.

Man: Thank you.

Woman: Thanks, bye.

(Andrew): Bye.

Woman: Bye.

END