# NATIONAL QUALITY FORUM

**March 22, 2019**
**7:00 am CT**

Poonam Bal: Hi everyone.  This is Poonam of NQF again, just letting you know that we'll be starting in a couple of minutes.  Thank you for your patience in the meantime.

Hi, everyone.  Thank you for your patience.  Welcome to the Second Subgroup Number 1 Call.  I'll be - while it was in the first one it is our last call of the subgroups.

So this is going to complete everything and then we can finalize list of what's going to be moving forward.  So as a reminder to everyone to panel members our discussion guide was sent out in that email.  Also a SurveyMonkey link was sent out.  Please make sure to keep that open and put your votes in as we go through the different measures to be discussed.

We will - staff will prompt you when it's appropriate to go ahead.  And we are discussing five renal measures that we were not able to get to at the last subgroup call.  And Andrew will be leading the discussion.  We might discuss one more measure if we have time for it but we'll see how we're going at that point.

With that, today’s call is limited.  We have two hours, unfortunately this is our last phone call.  We really need to get through all five measures today and make sure we’re able to come to consensus on anything that we have not been able to do so far.

And then also as a reminder, this is a public call.  The developer will be on the line to answer questions but there will be no opportunity for public comment. And also for record keeping purposes, please state your name each time that you are going to make a comment.  So it’s easier for us to know who is speaking and also to keep track of it later.

With that I’ll give it to (Andrew) to do role call and to start the discussion.

(Andrew): Thanks Poonam.  All right.  So I’ll just run through our subgroup reviewers here to see who’s on the call.   Bijan Borah?

Bijan Borah: Present.

(Andrew): Okay.

Bijan Borah: I’m here.

(Andrew): Great.  Thanks.  Paul Kurlansky?  John Bott?

John Bott: Yes.  Here.

(Andrew): All right.  (Jeff) Geppert?

Jeffrey Geppert: I’m here.

(Andrew): Great. Thanks. And (Sherrie Kaplan)? All right. Well, we can go ahead and get started and get our, get into our discussion here. So just as a reminder, we were reviewing on the last call a set of measures from our renal project. There were a number of concerns about the method of testing for reliability and the results.

And particularly I think we had some questions about the method for testing reliability around the bootstrapping approach. And we were hoping to have get some a bit of clarification from the developer on this call. There was also at least one measure where face validity only had been provided for a maintenance measure.

And we wanted to see if the developer could give us a rationale for only having face validity and try to make a determination on whether that was an acceptable justification or not.

So we can sort of just get started. I think maybe with, well, I don't know if we should go measure by measure or maybe we should allow the developer to make some remarks upfront and just get sort of a general overview of their stress testing approach if they're ready to do.

Paul Kurlansky: Hello?

(Andrew): Yes, hello?

Paul Kurlansky: Yes, hi. This is Paul Kurlansky. I'm sorry, I'm late.

(Andrew): Okay. Great no problem. We're just going to have our developers of the set of renal measures, give an overview of their testing methodology, particularly with respect to the bootstrapping approach or do we have the developers online?

Joseph Messana: Yes. Hi, Andrew. This is (Joe) Messana from U of M KECC.

(Andrew): Hi.

Joseph Messana: I'm here with our group, notably with Rajiv Saran, who's the other clinician on our group that worked on development of the KC overview measures and nPCR measures. And Professor Douglas Schaubel from our Biostatistics Department who spoke briefly on the last call.

And I think has more information about bootstrap the other reliability, the other IUR concern that had been raised, but not completely resolved. So I'll turn it over to Dr. Schaubel -- (Doug) -- now to speak.

Douglas Schaubel: Yes, so going back to the last call, one of the concerns was, well, one of the questions was, how many bootstraps resamples we did. So there were thousands bootstrap resamples. And the next question related to how exactly the bootstrapping was done. So it was done within facility.

So for each facility we actually resampled the individuals. And so we took the individual's experience over, say the calendar year and then resample back so that we're maintaining the correlation structure in the original data.

We'd also made some comments about normality and the issues surrounding that came up at the last call. And I think that's really a red herring. The idea is that if we're doing - if we're actually doing analysis of variance and if we're fitting a mixed model and trying to interpret the parameters of that model that does require normality.

And on the other hand the decomposition into between and within does not. Basically, we’re requiring that the mean exists, which is a very safe assumption in the context of these data and that the variance exists which is also a safe assumption.

So we can still meaningfully calculate between and - or rather within and total and take the difference to get between. And the ratio that’s taken to get the IUR is still meaningful whether or not we have normality. So, none of that depends on normality.

Again, if we were doing IUR differently. If we’re relying on a mix models to come up with those calculations that does require normality. But we were taking a pretty dramatic steps to avoid that. That’s where the bootstrapping came in.

So we did evaluate the skewness and the kurtosis for various measures and I could go through those though I think that’s really, that’s kind of a moot point at this stage. And I think our language was is a bit regrettable. I think that to say that the results have to be interpreted with caution because of potential non-normality I think I think that’s overstated. That’s all for now.

(Andrew): Okay thank you. Any thoughts or discussion on that from our reviewers?

Bijan Borah: All right. This is Bijan from Mayo Clinic. So I think one of the - I think this is (Shay) that who brought up - I was under the impression that some of these say (unintelligible) and other statistics would be shared with us before this meeting.

So I guess again, I mean, it's not that we're trying to kind of hold this whole thing, but it would have been really helpful if we would have seen those numbers right in front of us. But not knowing, I guess, you know, what was the sort of you know extent of (unintelligible) distribution of the data that you have and that would have been really helpful for us to see.

((Crosstalk))

(Andrew): Sorry, this is – yes, this is Andrew. Just to respond from NQF side. That was our determination just for the sake of consistency and I guess fairness with respect to other developers.

Bijan Borah: Okay.

(Andrew): We didn't really want to sort of allow for additional written materials. We felt that might be inconsistent again with, sort of, deadlines and requirements we'd have imposed on the rest of our developers. So we asked the developers in this instance to just give us an oral sort of explanation and report on the methods so.

Joseph Messana: And Andrew Dr. Schaubel is prepared to read them if it's required by the methods group. We don't want to withhold any information but for the sake of time I think his initial response suggests that the degree that with which he said are different from normal is not critically important here.

(Andrew): Okay.

Joseph Messana: Go ahead.

Douglas Schaubel: So in terms of skewness for adult Kt/V hemo so skewness is negative 7.3. It's negative 3.2 for PD - Adult PD Kt/V. For pediatrics hemo Kt/V it's negative 1. And it's also negative 1 for pediatric PD Kt/V.

Kurtosis for adult hemo Kt/V was about 91. For adult PD Kt/V was about 15. And for pediatric hemo Kt/V was 0.07. And for pediatric PD Kt/V was 0.14.

(Andrew): All right. Thank you for that. So maybe we can just, again have a bit of discussion over the general approach this has - is there or any thoughts from our panel members on the issue of the sort of normality or lack of normality and the distribution and are you sort of comfortable now with the reliability results are - reflect, you know, the true reliability of the measure or an appropriate approach to testing reliability?

Paul Kurlansky: All right. This is Paul Kurlansky. I had a different - completely different concern about reliability. I wasn't particularly bothered by this particular issue. So I don't know if it's appropriate to comment.

(Andrew): Well you can go ahead and talk about what your concerns were.

Paul Kurlansky: Well there were – okay there are number of measures but I'm just looking back at my review of 249 is that a bit…

(Andrew): Yes we can just kind of use that one to start with.

Paul Kurlansky: And I thought that the score-level testing was appropriate, but it didn't seem to be testing of data element reliability. And one of the references actually said, you know, no center - I'm just quoting. "No center apart from one unit managed to complete the collection blood specimen as recommended by the guidelines. One exception blood collection for hemodialysis adequacy was not performed using proper technique in any center." So this would suggest the data element used for the measure may not be reliable. And I was just concerned about the data element reliability testing would therefore need to be documented not so much the measure itself.

(Andrew): Did the developer have any thoughts or response on that?

Joseph Messana: Yes Andrew, we do. Is it time now?

(Andrew): Yes, yes, that would be fine.

Joseph Messana: All right. So I'd like to make sure before I try to respond to the reviewers concern that the meeting minutes should - needs to include a note that states that as a measure developer I'm really confused about how one study from an old evidence form submitted in 2015 that is not really being interpreted in the context or with regard to the studies limitations is potentially threatening re-endorsement of a familiar multiply endorsed measure. I'm really confused about that process because…

Paul Kurlansky: It's the first reference.

Joseph Messana: Well, let me…

Paul Kurlansky: You reference it, so…

Joseph Messana: I did not reference it. It was referenced in the 2015 evidence form submission which we have not yet submitted, which is not on our current list.

Paul Kurlansky: I don't know where I got it from then. I got it from that was submitted.

Joseph Messana: I don't either, which is part of the problem from my perspective. Now let me as what I said…

Paul Kurlansky: It says (unintelligible)…

Joseph Messana: I understand. I'm about - I'll discuss the article if you'd give me just a minute here. Prior NQF Renal Standing Committees made up of individuals with extensive dialysis expertise have not expressed concern about this issue. The dependency of accurate Kt/V on appropriate blood draw technique has been known for over 20 years.

And the dialysis communities comfort level with it has been demonstrated by this measures repeated endorsement by NQF over the last decade. The (Ulusoy) article that you refer to include it in the 2000, I think, 15 submission, describes a clinical study from nearly a decade ago in eight Turkish dialysis centers. It is not generalizable the US dialysis centers and has significant methodological limitations. It's not well blinded. It's not a crossover study, so the study effect bias is very likely.

The dependence of Kt/V on appropriate blood drawing at the end of treatment has been a known issue in the dialysis community for over two decades. Appropriate techniques regarding blood drawing technique were included as part of the 2000 KDOQI hemodialysis advocacy guidelines reinforced in the 2008 update to the ESRD conditions were covered by Medicare with a specific V-Tag citation for failure to use proper technique. And it's in the Medicare claims processing manual revisions from 2010 when Kt/V was added as a required data element. This is old news.

That study, I don't know why that study was included in the evidence forum in 2015. But probably to demonstrate that appropriate technique is needed.

But regulatory and standard practice in the dialysis community is known about this - and the US dialysis communities known about this for probably 25 years.

There is literature at least one single center study that demonstrates that the mismatch between prescribed from urea kinetic modeling and delivered which would be the result from inappropriate blood drawing at the end of the dialysis as the main reason.

About 3% of samples show that in the 2003 study from the Case Western group that was published looking at about 800 monthly dialysis specimens. There is no widespread evidence that this is operative in the US dialysis community. And the renal standing committees' endorsement of this measure in - for at least two prior go arounds is strong corroborative evidence that they're not worried about it. Thank you.

(Andrew): Thank you.

Paul Kurlansky: It would help if that evidence were included in this admission.

Poonam Bal: So at this point, we don't require evidence to be included in the review. I think there might have - you know, we do try to include as much information as possible for your review. But then also I think if you - in the study (unintelligible) that contradicts whatever is found. If you were to - find that you should be using that to your - for your knowledge. But thank you for your response tool. (Andrew) I'll give it back to you to continue on.

(Andrew): Yes, appreciate that. And yes we - I should note that the developers will have an opportunity to update their evidence form. That's something that we're, you know, that they'll - they have a bit of time to do before the official submission date. We're sort of still supposed to be reviewing just the specifications and testing information at this stage.

Is there any thoughts or comments from our reviewers on that explanation or sort of response to the concerns about data element?

I don't know if - I just want to check it. Has (Sherrie) joined the line yet? Doesn't sound like it. Okay.

Jeffrey Geppert: This is (Jeff). I have a couple other questions just…

(Andrew): Yes, sure. Go ahead (Jeff).

Jeffrey Geppert: Yes, so I am not challenging the developer in any way I'm just - so you don't need to get defensive but I'm just trying to understand just for my own knowledge. This is the first time I've actually seen the IUR used as a metric of reliability. So can you just briefly explain sort of the rationale for choosing that particular metric, you know, versus the things that are more commonly used, like, the Adams approach or…

Joseph Messana: So yes, this is (Joe) Messana. I'm sitting around the table with two senior biostatisticians here at University of Michigan. And we've been submitting measures to the National Quality Forum for over a decade since I think 2007. And we were asked to start using IUR for reliability. I believe it was under 2011 by the National Quality Forum. And so I'm confused by the question.

Jeffrey Geppert: Like I said, you don't need to get defensive. You're very defensive today. I understand, you know, this is sort of…

((Crosstalk))

Jeffrey Geppert: … review process. And I'm just asking. So you're explaining something to me…

Joseph Messana: If you sense…

Jeffrey Geppert: …that you were told - that you were told by NQF staff that that was a good metric to use…

Joseph Messana: Yes.

Jeffrey Geppert: …so thank you for that explanation.

Joseph Messana: If you sense defensiveness, it's because we believe very strongly that these measures move forward. And we were concerned about the possibility that they would not.

Jeffrey Geppert: Okay. Yes, no, I mean, it's good that you're passionate about your measures. And we're just trying to, you know, use due diligence in terms of our review and make sure that we understand. And I think, you know, you've done a good job of sort of explaining your thought process. And so I think your point about sort of being - so we'll have to look into that. Sort of talked to NQF about that.

The other sort of, again, sort of informational question I have is there was an article last year and maybe you're actually on the call. Or - is one of you Kalbfleisch - (John) Kalbfleisch?

Joseph Messana: Dr. Kalbfleisch is in the room yes.

Jeffrey Geppert: Okay. So you wrote an article last year in Health Services Research about the use of IUR as a measure of reliability where you advocated against it. I - my interpretation was. So I just - could you just kind of explain sort of that article and kind of the discussion today?

Jack Kalbfleisch: Yes, so I guess it was reacting a bit to the very strong recommendations that we've got from NQF from the past that reliability should be measured entirely by IUR. And I think sometimes that's appropriate and sometimes it's not. But there's really been no distinguishing like that. So it's just it's been recommendations of IUR has been that we should be reporting IUR on every measure. And they even put in the lower bounds for it which I think are not particularly sensible.

So it's something that we've had some discussion with NQF in the past on occasion. That article was written I guess partly to discuss a bit what the rationale for IUR and sort of measure of reliability is. This is under which it would make sense to stay with that and the conditions under which you should really relax that interpretation.

So in the sense IUR really - as it's proposed in the Adams manuscript that I guess NQF now refers to is really an appropriate measure of reliability if all the inter-facility or inter-hospitals differences are due to quality of care.

So as long as there's no unmeasured confounders that are affecting the facility levels and it's an appropriate measure. And I think that's rarely the case. So it still tells you something about signal to noise. But I think as a specific measure reliability, it's not generally the best thing to do.

Jeffrey Geppert: Okay good. Thank you very much. And just my last question is about sort of the risk adjustment. So you don't risk adjust. It's an intermediate clinical outcome measure. That’s common. I mean, most intermediate clinical outcome measures are not risk adjusted. But can you just sort of speak to that? Is that - I mean, do you have concerns about that or are you comfortable with that? Is the community comfortable with that?

Joseph Messana: Trying not to be defensive. This is (Joe) Messana again.

Jeffrey Geppert: Yes.

Joseph Messana: The community has been comfortable with not risk adjusting these measures as evidenced by prior recommendation for endorsement by the Renal Standing Committee.

These measures are really fundamental basic outcomes that are minimum standards kind of really more focused on exceeding that dialysis facilities deliver in excess of minimum for both for the (unintelligible) clearance and for the avoiding hypercalcemia measure. And they have good success at achieving these results across the board. But they are really exceeding. These measures were designed to be consistent with regulations and minimum requirements recommended by the consensus endorsement documents.

And so we don’t think they need to be risk adjusted because most dialysis patients should be able to achieve these values. And so risk adjustment has not been deemed necessary. And I think the dialysis community generally agrees with that.

Jeffrey Geppert: Okay. Good. Thank you. Thank you very much. Appreciate your answers to the questions. That’s all I have.

(Andrew): All right. Thank you very much. So was there anything else that anybody wanted to discuss about Number 249? I think we've gotten some responses about the method and approach for assessing reliability. Talked a little bit about the risk adjustment and seeing if there's anything else we need to talk about.

I think those were the major issues that we had identified in our preliminary reviews. Any other discussion of 249 or shall we go ahead and take a revote?

John Bott: Yes this is John Bott. A comment is that relates to a number of these intermediate outcome measures and as far as what (Jeff) Geppert just asked about and as far as risk adjustment goes. It seems the response oftentimes and with this measure and a number of other ones is inadequate. The measure steward/developer stated something to the effect of risk adjustment is not necessary.

But we really don't have any evidence presented to us that it wasn't necessary. And so they don't go on to say, you know, we checked. Here's what we found. Thus, it's, you know, we don't think it needs to be risk adjusted. It's like, we just have a lack of adequately defending the lack of risk adjustment and I thought for outcome measures and intermediate outcome measure -- correct me if I'm wrong -- it - you know, you need to state your case more empirically why you're not risk adjusting. And I just thought there - a concerted effort should have been made to vet just to see if it would have benefitted by it and I didn't see that. So I thought there needed to be a better case made to defend the lack of risk adjustment.

(Andrew): So we have sort of established that expectation for outcome measures. I don't know that we've sort of historically communicated that kind of expectation to developers for intermediate outcome measures. It is something that we're talking about now as sort of the approach to, you know, reviewing risk adjustment with intermediate outcomes.

It's and I should also note and -- (Karen) correct me if I'm wrong about this -- we're trying to sort of guidance to the methods panel has been that we shouldn't stop measures at this point due to a lack of risk adjustment or, you know, any particular concerns about risk adjustment approach because that's something that we're sort of seeing as more in our standing committee's domain with their clinical expertise in the topic.

So we can certainly communicate concerns about a lack of risk adjustment to the Standing Committee along with the message panels, other findings and ratings. But it's not something that in itself should hold up the measure at this stage of review. Is that correct, Karen?

(Karen): This is (Karen). Not quite (Andrew) but you're close. So going back to you, well first of all, let me just what we asked the methods panel not to do is not to fail a measure simply because they disagree with the inclusion or lack of the inclusion of certain factors and the risk adjustment model. And that usually that has come up primarily in social demographic types of factors.

So - but it's still certainly within your purview to settle whether or not risk adjustment, you know, you feel that risk adjustment overall should have been done or, you know, that the risk model calibration discrimination that sort of thing is reasonable. So that is within your purview but we ask you not to quibble and fail on it, you know, specific factor or two or something like that.

Going back to John's question, we would love to see empirical analysis demonstrating that risk adjustment is not needed. However, for some measures and usually it's outcome measures. But I think I would probably treat intermediate outcome measures to, you know, in the same way. We could if it's reasonable for you guys, we have allowed things to go through with a conceptual rationale.

You know, and I'll give you an example. You know, sometimes people will say, well, in hospital mortality, you know, we don't need to adjust for certain things or maybe we don't need to adjust. And that might not be a good example and I apologize if that's a bad sample. But, so John, you could personally, we will convey that you would like to see analysis, you know, really kind of demonstrating no need for risk adjustments. But at minimum, we would need to see some kind of a conceptual narrative about why they didn't do it or why they don't feel that it's needed.

I hope that's clear. John, is that clear? And (Andrew) did I? Would like to know.

John Bott: Yes, this is John. I understand what you're saying. The instruction on our PA form whatever you call it for Question 16 in the comment area, it says Question 16 which risk adjustment it applies to all outcome costs, et cetera, et cetera measures.

So what I thought you meant by outcomes includes, you know, pro PMs and intermediate outcomes unless you specify otherwise. So you might want to think about how you - particularly the outcomes when you said the outcome.

(Karen): Yes, no. It definitely includes intermediate outcomes per PMs, just regular health outcomes. So, we definitely want you guys to be thinking about risk adjustment and need for. But all I'm saying is, if a developer has said, you know, they provided some kind of a narrative to describe conceptually, you might be willing to, you know, we're okay with leaving it that and not having actual empirical results.

But that's something that we could talk about in the future. You know, should we always require empirical and, you know, not just the conceptual side.

Joseph Messana: This is (Joe) Messana at KECC if we would offer a comment from Dr. Kalbfleisch related to this matter if you believe it's appropriate at this point.

(Karen): This is along the lines of more of a conceptual narrative is that what you're thinking?

Joseph Messana: I never know exactly what Jack is going to say but I guess so. He is the most that means of it.

Jack Kalbfleisch: At this point, I was just wanting to make and I think it's one that (Joe) made too but it's just that what this is a process measure. And it's sort of a level of care that one expects of every individual. And that measure is really checking on that. And you wouldn't really want to adjust that for age, I mean, because it's not like that measure should be different for younger people or older people. They all should get that level of care. And a measure like that, really risk adjustment doesn't make that much sense. I think where it make sense is where your outcome is something like mortality or hospitalization or readmission where there are aspects of that that aren't under the control of the facility.

And consequently you don't want to penalize some because they haven't - because they've got old patients that are dying quickly or other things you want to take that into account. That's not the case with a process measure where you really want to deliver a standard of care to everybody. And I think that's the difference.

((Crosstalk))

Paul Kurlansky: This is an intermediate outcome measure, and now you're calling it a process measure. So it's essentially defined what or how you're defining your measures.

Jeffrey Geppert: Yes this is (Jeff). I might take on that, I think I sort of just mentioned this last time is that when - so it's labeled as an intermediate outcome measure I guess because of it's a clinical - it's the health status is measured by a lab to value. You know, but it's a health status that is sort of completely determined by a process of care.

So to John's point conceptually it is like a process measure, even though. So it's kind of this weird sort of hybrid, you know, that I think is generating a little bit of confusion. But I think the explanation, conceptual rationale sort of made sense.

Paul Kurlansky: I mean, I guess the conceptual rationale is that there's no valid reason to believe that any of those - any risk factors would impair your ability to deliver the service differentially.

Jeffrey Geppert: That's a little different issue. You know, that's again, it's sort of like, yes, like this is the sort of the quality construct varied by patient characteristics because it's harder to do this for some people than it is for others that might be sort of a case mixed adjustment issue.

Risk adjustment really has more to go it’s been like, you know, is it that, you know, is the likelihood that you’re going to actually achieve the clinical health status. You know, vary based on patient characteristic regardless independent of anything a provider might do. And I think what you’re saying is that that’s not true.

Paul Kurlansky: Right. That’s my understanding.

Man: Right. Yes.

Jeffrey Geppert: Yes, yes so that makes sense. It’s a mechanistic kind of relationship.

(Andrew): All right. So any additional discussion here? Or again are we ready to take a vote? Without objection I think we will go ahead - sorry. We’ll ask you to go ahead and fill out your SurveyMonkey. We are only voting on reliability for this measure. So just pass on.

Man: And just repeat again with the measure number, is it (1)

(Andrew): Sorry yes. Thank you, 0249 is the minute measure you’re voting on right now. So just to vote on reliability.

Paul Kurlansky: Could you send the SurveyMonkey again?

(Andrew): Yes, we would send it right now. In the meantime, we can probably move on to the next, which I believe has very similar issues here. Let’s see I’m just pulling it up, Number 318 is it…

((Crosstalk))

Jeffrey Geppert: And just to be clear, so we’re not be voting on the validity because we voted on that last time?

(Andrew): Yes, correct, just reliability.

Jeffrey Geppert: All right thank you.

(Andrew): And I think we’ve got the same basic issues for this measure as we did the last one. So does anybody want to continue the discussion or are we comfortable?

Paul Kurlansky: Which - I’m sorry which measure are you looking at now?

(Andrew): Number 0318, delivery dosage of peritoneal dialysis above minimum?

Paul Kurlansky: And our concern is with reliability?

(Andrew): Yes, there was that we got a consensus not reached rating on reliability. Same kinds of concerns here that they wanted - the reviewers wanted additional details on the bootstrapping method. Some concern about the method of assessing reliability with the IUR approach. The same questions about data element reliability.

Let’s see. And then just for this one, there was some question - there were some questions about the specifications. And some reviewers were unclear about certain aspects of the numerator. The numerator details include a note that missing, expired and not performed are not counted as achieving the minimum weekly Kt/V threshold. And reviewers were not quite clear on whether these instances were counted as a 0 in the numerator or treated as exclusions or exceptions, so to speak.

And maybe that's something we can get clarity from the developer on. My impression was that it does appear to be treated as a 0 in the numerator. Is that correct?

Joseph Messana: Yes, that is correct. As long as there is not a Kt/V value within in the prior three months.

(Andrew): Right.

Joseph Messana: So that is correct and that specification is consistent with the consensus endorsement statements about peritoneal dialysis Kt/V. Unlike hemodialysis, it's required every four months, not every month, according to the consensus of statements.

(Andrew): Thank you.

Jeffrey Geppert: Yes this is (Jeff). Can I (unintelligible) process question?

(Andrew): Sure.

Jeffrey Geppert: So (Sherrie) wasn't able to join us today.

(Andrew): Yes.

Jeffrey Geppert: Is she going to have a chance to weigh in? Can she listen to a transcript or something?

(Andrew): Yes, we'll send her the recording and ask her to listen to it before she - or actually she may not have the opportunity to vote on these since she missed the call. So, we may…

Poonam Bal: So yes, we hit quorum on this call, meaning we have enough participants to go ahead and vote. So we will share the transcript and the recording with her as well to public but she wouldn't be re-voting on it. Only the people on the call would just make sure that it's consistent.

Jeffrey Geppert: Does her past vote apply?

John Bott: This is John Bott are we voting now on 318 or are we about to begin discussion of 318?

(Andrew): Well, that's the way we're seeing if there is any discussion. Do you have any comments or questions before we go to a vote?

Again, we got some sort of response at the beginning of the call on the - in regard to the bootstrapping approach and the Inter-Unit Reliability testing method. I don't know if there were any additional questions about that. We did hear - (Karen) reminds me that it may have been (Sherrie) who was asking this. I think she had been asking about how many patients or people were included in the bootstrapping approach. I think maybe just one more reiteration from our developer might help. Again, how many, you know, resamples were taken, how many patients were included that sort of thing. Does that make sense?

Douglas Schaubel: Yes, this is Douglas Schaubel again from KECC. There were a thousand bootstrap samples taken. We bootstrapped at the individual patient level. So all observations within a year. And the bootstrap was done within facility.

((Crosstalk))

Michael Abrams: Can you just clarify -- Michael from NQF here -- when you bootstrapped how much of the sample did you take? Does that make sense? So you did it 1000 times, we get that, right? But how many did you take? Did you take half of it? Did you take…

Douglas Schaubel: So the data - there's not a reason for doing less than the entire data set or less than the size of the entire data set unless you had maybe 2 million patients and it was computationally overwhelming to do so. Then you - you had that discussed last call. You could take a fraction and then scale the results down. But we didn't do that.

Michael Abrams: You did that with the whole thing every time is that what you're saying? I'm a little mystified…

Douglas Schaubel: Yes.

Michael Abrams: …by that. But is that - that's the process?

Douglas Schaubel: Maybe I would have been computationally challenging 15 years ago but it wasn't these days.

Paul Kurlansky: I guess it just doesn't sound like bootstrapping. Doesn't like…

Michael Abrams: To me neither.

Paul Kurlansky: It's not a sample.

Michael Abrams: Got to be a sample.

Paul Kurlansky: That's what bootstrapping is.

Douglas Schaubel: Maybe you have a different idea what the bootstrap is than me. And then this the (unintelligible) statistics.

Paul Kurlansky: I mean, maybe I'm wrong. I thought bootstrapping should take us a random sample multiple times and then you see…

Bijan Borah: But did you - this is Bijan…

((Crosstalk))

Paul Kurlansky: …certain things emerge.

Douglas Schaubel: If you - so then let me address this differently. So suppose - this is the way it was done previously, many years ago. So if you took your original study population and magnified it. So in other words created a super population then if it's from your original it would have the same properties as your original dataset. And then you could take samples from that super population. And unless you had a good reason for doing something different both samples should be still the same size as the original facility sizes across the country.

So, if that's what you're thinking of - you - what we did is equivalent to that. Just technically it's different in terms of the code that went into the process, it's different. But conceptually it's the same thing. If I'm understanding what you're trying to say, you think we should have done.

Bijan Borah: So hi this is Bijan. I think you did it what I understand as bootstrapping. So just give you a complete example, say facility A has say, you know, 100 patients for example. So in each of your thousand bootstrap applications you essentially (unintelligible) from that 100 patients essentially make a draw of 100 patients with replacement, correct? I mean that's what typically you do.

Douglas Schaubel: That is what we did.

Bijan Borah: And then you…

Douglas Schaubel: That's what we did.

Bijan Borah: Okay (unintelligible). Yes. Okay.

Michael Abrams: Again, Michael from NQF here I'm not a statistician, right? I've had some training in it, but is the committee comfortable with how the bootstrapping has been described? And if not, please do seek clarification now if you can.

Bijan Borah: This is Bijan. I think I understood, I am good.

Jeffrey Goppert: Yes if he's good, I'm good.

Paul Kurlansky: (Unintelligible) with that but okay.

(Andrew): All right. And that being said, unless there are any other issues that you'd like to discuss.

Paul Kurlansky: I had a very minor point about the specification about the reliability.

(Andrew): Sure.

Paul Kurlansky: And that is did we know that or CMS know specifically that the measurement has been taken? In other words how do we know that it's not the measurement from three months ago that’s just being, you know, re reported and then the same one is re-reported? Did we actually - is there a way to know that - is there, you know, with your claim file when the measurement is taken? Is there some way to know that a new and current, you know, measure at least within the last four months measure has been taken?

Joseph Messana: Yes, so this is (Joe) Messana. So both CROWNWeb data sources and claims data sources for this. I have a date the test was performed, every time it’s performed. And we in our database, we maintain the date along with the Kt/V determination that’s reported by the facility either in the CROWNWeb or claims. That’s how we use the current months or prior three months algorithm to determine missingness or presence. So, all of those are store, yes.

Paul Kurlansky: Perfect. Yes. You know, it wasn’t clear to me. So, thank you.

(Andrew): Okay. Great. Thank you. So, are we ready to vote on Number 0318? If no objections, well we’ll ask you to go ahead and do that. Again we’re voting only on reliability for this one.

Okay. So, next we’ll go on to the measure…

Bijan Borah: (Andrew)?

(Andrew): Yes.

Bijan Borah: A quick question. Have you changed or sort of modified the SurveyMonkey form? Because like the earlier time, now we get back to the new survey form right here. And earlier, like, you know, once we get it up then it just goes up and then we have to link - we now click the link again. So now, which is a good thing. But is that something you did or so am I getting it wrong?

Poonam Bal: To be honest, the person that usually handle SurveyMonkey is out today and I'm just going to give her credit for fixing it.

Bijan Borah: Okay. All right. Thank you.

Jeffrey Geppert: As long as you're not like rewriting our old responses over and over again.

((Crosstalk))

Man: Okay.

(Andrew): All right. So yes, please enter your vote for reliability on Number 0318. And we will move along to measure Number 1423, 1423 is minimum STKTV for pediatric hemodialysis patients.

Again, some of the same, similar kinds of concerns here that we've already talked about on the bootstrapping method, IUR, the missing values. I think the - so I don't know that we need to talk about reliability again unless does anybody want to say anything more about reliability on this issue - or on this measure?

Paul Kurlansky: I'd sort of, I think it's been answered by the response to the first question I asked. But when I saw this, this one is a pediatric one?

(Andrew): Yes.

Paul Kurlansky: Yes, no. I guess because there are fewer centers and whatnot I was more concerned about the limited number of available centers if there was center reliability or that had been tested if there was evidence regarding that in the past. It may have been asked and answered elsewhere. I just didn’t see it in the information that I had.

In other words is there any variability in the way this is measured from the center to center?

Joseph Messana: This is (Joe) Messana. I would offer the same responses to your question for the adult. They use a standard approach, pre dialysis, post dialysis BUN with appropriate blood sampling technique. And the calculations used are the same. The target recommendation is slightly higher based on the pediatric nephrology communities’ recommendations when the endorsements were put together.

Paul Kurlansky: All right. Okay. Yes, that’s what I say. I think you answered this question. This was a question I have when I read the thing but you answered it when you answered my first question. Okay, so I’m good.

Jeffrey Geppert: Yes, I think the only difference here is in the adults, there were several, they’re like 1400 facilities, something like that.

Paul Kurlansky: Right. That’s what raised the question in my mind is whether or not there’s variability in the kids.

Jeffrey Geppert: Now there’s 14. Yes. So, do you have any concerns about the fact that they’re only 14 facilities?

Paul Kurlansky: I think if it’s that will standardize, I don’t think it should matter.

(Andrew): Okay. All right. Well, go ahead.

Jeffery Geppert: So there's over to sort of, again sort of just comment on the, this is a topic that we've had a lot of internal discussion about is the count, you know, so the idea of sort of minimum thresholds, and you're, you know, pretty consistent that you require at least 11 cases in your testing. Is that also the way it's reported? So, you don't report any data…

Joseph Messana: Yes, that was correct. And the reasons I think are historical but a significant part of that threshold less of not reporting less than 11 is because of potentially identifiable individual cells. You have a dialysis facility with a location and reporting 8 or 10 is seen as potentially sharing identifiable information. And so we did the testing using the same approach that we do for public reporting.

Jeffrey Geppert: Okay, great. Thank you.

(Andrew): All right. Thanks. Unless there are any objections, we'll move on to validity. Any more comments or questions on liability for this measure? If not, we can go ahead to validity.

We did have - for this measure which is a maintenance measure was submitted with face validity only. And so we wanted to hear from the developer what their justification for only giving face validity for this measure. So we can maybe hear from him now on that subject.

Joseph Messana: So this is Rajiv Saran Professor of Internal Medicine, who led the (unintelligible) group to respond.

Rajiv Saran: A long time ago but…

Joseph Messana: Yes.

Rajiv Saran: So recalling from 2010 I mean, we can go back to the transcripts. But the main issue that was discussed around the table at the time was that there's no actual data in the pediatric literature to support development. Kt/V was developed as a metric in adult population, pediatric publishers are not specifically studied however - yes.

Joseph Messana: And so I think the - no the specific question now is (unintelligible) for reendorcement of face validity that we have not accepted without empiric testing. And so results of empiric testing are what they're asking for.

Rajiv Saran: So that is no - they go by the same threshold and guidelines that are used for the adult population. The argument that was made for face validity is the fact that in population achieving these targets is even thought to be of greater importance because of kids' growth concerns. The concern was nutrition and growth for kids. And they said unless we endorse minimal thresholds, we may be doing the community as a disservice not having any standard at all.

Joseph Messana: And I thank you. What I would add is that for this submission, there is very small number of facilities would not base on preliminary calculations would not have shown any associations. I think the limit here is sample size.

(Andrew): Okay. So that is the rationale presented for face validity. Any discussion of that by our methods panel reviewers?

Paul Kurlansky: I guess, I mean, I sort of accept the face validity but I was a little, I guess disappointed because it's not a new metric as you point out. And so in the last 10 years I would have hoped - even it's only 14 sites that would have hoped that they would be a collection of - an effort to collect empiric data to validate to measure that you - I understand the rational you've adopted but on the other hand it was sort of a minimal rationale. You actually have reason to believe that perhaps for kids maybe should even be higher than the adult standard.

And so with that concern it would seem maybe ideal but certainly better if there were empiric data to validate the threshold that you've chosen.

Joseph Messana: I appreciate that. The empiric results - identification of statistically significant associations with primary outcomes are negative essentially because of very small sample size.

Rajiv Saran: If I may add -- this is Rajiv Saran -- the duration of follow-up in the pediatric communities is often shorter because they get - they don't stay on dialysis for long enough to observe outcomes like mortality, especially because…

Joseph Messana: Early transplantation…

Rajiv Saran: …early transplantation is usually the norm. So that there's not enough.

Man: Thank you that's (unintelligible).

(Andrew): Okay. Any other thoughts or comments or questions for our developers or are there points of discussion you'd like to raise for our reviewers?

Jeffrey Geppert: So just as a thought experiment, so I know that, you know, perhaps there aren't many facilities and there aren't that many kids. And with kids, it's always hard to do outcomes because they don't, you know, they don't like you said, I mean, mortality is not sort of relevant. You mentioned transplant as a sort of progression. If they don't transplant, then what happens to them? They just stay on dialysis?

Joseph Messana: Is that a question for the developer or…

Jeffrey Geppert: It's a question is for the developer, yes.

Joseph Messana: Okay. Well, so that is…

Jeffrey Geppert: My basic question is if you were to conceive of a validation study, what would it look like?

Joseph Messana: So - well the outcomes that have has been shown to be affected by adequacy of dialysis since we're talking about the pediatric Kt/V going back to the original study looking that for (unintelligible) kinetics arose in adults have been hospitalization. And particularly hospitalization for uremic symptoms and mortality, right?

And so most of the observational studies link to those two primary outcomes. So the person's transplantability could in theory be affected or influenced by inadequate dialysis because of general health deterioration, poor nutrition or whatnot. But that would be less direct, I believe, than what has been shown in the literature, predominantly for adults as Rajiv has pointed out, which is hospitalization and survival - or hospitalization mortality.

Jeffrey Geppert: Okay. So if it is hospitalization there's an endpoint for a validation study even if you don't think the results would be significant, but?

Joseph Messana: Yes, yes, exactly.

Jeffrey Geppert: Okay.

(Andrew): All right. Are we ready to vote on 1423?

Jeffery Geppert: Does NQF have any sort of general comment on this issue? So we expect empirical validity testing for maintenance measures. This was a maintenance measure. It's actually been in the field, you know, for a decade. Developers provided some reasonable rationales for why doing validities would be difficult. You know, small number of cases and, you know, so what? So, we could say in this case, we could say we accept that rationale and sort of passed it on validity based on I guess reverting it back to the face validity, which is almost a decade old?

(Andrew): We do allow that...

Jeffrey Geppert: So I guess - go ahead.

(Andrew): Again we expect - we do have a sort of general expectation of empiric validity testing on maintenance review if - they are allowed to present face validity if they can provide a rationale that is acceptable to the, you know, group that is reviewing it.

So that's sort of your judgment call to make at this point whether they've given you sufficient or, you know, convincing rationale or not.

Jeffrey Geppert: Yes, all right. But there's assuming it sort of passes at this stage and successful with the steering committee, you know, it's three years between now and the next maintenance review and I guess part of the feedback would be to try really hard to think of, you know, a way to sort of assess the validity of this measure.

(Andrew) And that's certainly feedback we can provide to the developer and they're hearing it right now as well. All right, well…

(Karen): This is (Karen), sorry and - you know, I was thinking about this, I think, you know, sometimes, you know, whether or not you accepted justification of, you know, no empirical testing for maintenance measures sometimes, you know, that might be good to have purely kind of data-driven question. But other times, it might be a little bit more along the lines of some clinical realities. And in that case, I think that the message panel, you know, would err on the side of accepting and, you know, allowing the Standing Committee to make that final determination.

Man: Okay that makes sense.

John Bott: Well, this is John Bott my understanding is the Standing Committee makes the final determination anyways. But our role is just advisory, right?

(Karen): Yes. Well, it's advisory but that knowing that you're very steep in the methods so yes they make the - they could certainly vote differently than how you guys do. So yes sure, correct.

Poonam Bal: And if you were to hypothetically not move this measure forward, it would not go to the Standing Committees they would not get to make that decision. So you are kind of a gatekeeper. There is some decision-making on your part in terms of the measurable move forward to the Standing Committee but they will obviously make the ultimate endorsement decision.

(Andrew): All right, well, unless there's any additional discussion or any objections to doing so we can move on to a vote on Number 1423. For this one, we will vote on both reliability and validity. So, go ahead to your SurveyMonkey link and it's a rating for reliability and validity for Number 1423.

And maybe we do have - so we've got a couple more measures, maybe we could actually skip over 1454 for right now and go to Number 2706 just because this is a pediatric peritoneal dialysis adequacy achievement is target Kt/V.

And I think this is very similar to the issues that were raised on the last measures so maybe we can kind of just see if we have any additional discussion or if we can similarly just take a vote on reliability and validity for this measure as well.

Bijan Borah: And sorry what was the number?

(Andrew): Sorry, 2706. And for this one, we actually reached a low insufficient rating for reliability but just because the issues raised again were very - essentially the same as the other ones that we had a consensus not reached decision on. We thought it would be reasonable to go ahead and revote on this one as well, just to make sure we were consistent across.

Again, sort of reiterate here the, you know, we had an inter-unit reliability fairness with assessing with the bootstrap approach. It says it's the same kind of issues we've just been discussing. Then again, we have a face validity provided and I assume the same rationale given that this is also a pediatric measure and taking a look here. Yes, that's pretty much the same issues we just discussed.

So any additional comments or discussion on measure 2706 or are we okay, going ahead and voting on this one?

John Bott: Just to, maybe the developer maybe just speak to the denominator exclusions, my, I'm just trying to, my recollection was that it was a testing document.

There wasn't' - there was sort of a statement that there weren't any exclusions. Go ahead.

(Andrew): Yes, this is (Andrew). There was a statement that there were no exclusions but they mentioned in that statement, a number of exclusions that are implicit in the denominator. Basically, the way you're defining your denominator, population sort of implies, you know, exclusion of some other patients. I don't know that so we - we wouldn't consider that, you know, it's sort of the difficult kinds of exclusions. It's really just a matter of defining the denominator population, if that makes sense.

John Bott: That would make sense of the exclusions were all sort of clinical. You know, like, you have to have ESRD in order to be the denominator, right? That's, there're some of them and I don't really like that. It's like, patients who haven't been assigned for a month, you know, patients, you know, I mean, so fine but I mean but it says that's that kind of exclusion isn't really, you know, inherent in the denominator definition.

Joseph Messana: If I may, this is (Joe) Messana and I agree fully, it is confusing and we apologize. I think when this was written up originally the intent was to try to be as explicit as possible. And sometimes when you do that you create more confusion.

The denominator definitions excluded adults, excluded patients who hadn't been under the care of a facility for a minimum period of time, et cetera. So they're not really clinical exclusions as the way the term is used and it's unfortunate that we wrote it up that way.

John Bott: Okay, thank you.

(Andrew): Okay, thank you. Any additional points of discussion on this measure other concerns or points of clarification and thing like that?

Paul Kurlansky: No, my concerns were all things that have been discussed.

(Andrew): Okay, if there's no objection we'll go ahead and take a revote on reliability and validity for Number 2706. And then I think we can go back and jump up back up to measure Number 1454 proportion of patients with hyper.

We had a low insufficient rating, preliminary rating for reliability but again, had the same kinds of concerns we had about the other measures that were consensus not reached.

So I thought it would be, you know, reasonable to discuss and revote on this one again, just for the sake of consistency. I should also note that the developer did actually include some additional reliability analysis for this particular measure by doing some facility level, peers and correlation coefficient between the current performance month and the preceding month that and got, you know, a fairly strong results on that as well.

So I just wanted to see if there were any additional comments on reliability for this measure?

Paul Kurlansky: All right. I have questions about - I'm trying to remember the measurement. I had questions on the specifications.

(Andrew): Okay.

Paul Kurlansky: There were a bunch of questions that I had here, just sort of, I'm not sure I can summarize them, maybe we'll just go through them one by one. It wasn't clear to me if there was any standard regarding how frequently the calcium levels need to be drawn.

Joseph Messana: So this is (Joe) Messana, implicit in the measure definition since it’s basically reporting of patient quarter, so three month periods of time. And the measure specifies that you take the average of the monthly reported calcium values in that quarter for any given reporting month so then it rolls to the next month.

Soto be flagged as a zero in the numerator, you, for missing this, you have to not report a calcium value for three months, three consecutive months. So if you have less than three values in that reporting quarter, you take the average of whatever values you have one or two, obviously two values or you - or you take the one reported value.

So the minimum requirement is a calcium every three months which again falls within the generally accepted standards of care in the dialysis community and comports with the two tabs that reviewed this as being kind of reasonable appropriate.

Paul Kurlansky: So and, you know, going back to our previous discussion obviously CMS is going to know when you drew calcium levels but if you, is there anything to prevent somebody from gaming the system?

In other words, let’s say you have in the three months periods, you have three calcium levels and only one of them has really falls within the acceptable range. So and you report that one. Is there a way that this metric will know that you’re not reporting to which have been done?

Joseph Messana: No.

Paul Kurlansky: This is just about gaining the system that's all.

Joseph Messana: I appreciate that. I guess the - and I understand where you're coming at from that point of view, I believe that that's why the original tap I was in the room and when the tap develop this and then subsequently when the 2013 tap endorsed keeping it as is. That's why they relied on three months. It's a pretty low bar actually.

So if someone has three calcium values in the month, two of them are elevated, one of them is normal. The tap would say well, you know, we're not clear that there's a safety signal here. Its persistent elevated calcium was what they were driving at. And so if it's - if you have that much variability or lack of consistency in the hypercalcemia definition, they would say the measure, I believe they would say the measure is behaving as designed.

But I understand your concerns. I would be at least glad that the dialysis facility was paying that much attention to the patient serum calcium to measure it repeatedly in a month.

Paul Kurlansky: And this is, like, really very technical but is there any criteria that say - that state that the center has to use a CLIA certified lab? I guess he does - I mean it does in order to report to CMS is that…

Joseph Messana: Yes. So for Medicare patients and dialysis facilities fall under CLIA certification having been cited in my own facility for a glucose monitor that wasn't meeting testing. So most dialysis facilities now send their labs out to CLIA certified labs that are owned within a corporation. But, you know, same requirements as for any treatment of any other Medicare patient.

Paul Kurlansky: Got it.

(Andrew): All right, any additional thoughts or comments on Measure Number 1454?

Jeffrey Geppert: This is (Jeff). Just a question about the sort of analysis on statistically meaningful differences in performance. So it's helpful that they sort of did a test and then they sort of repeated the test with more recent data. So in the earlier results, I guess they're about 14.9% of facilities that had a performance worse than expected. And then in the most recent submission that's gone down to 7.3% of facilities what's the end game here? You know, like, at what point, are we trying to get to zero? What's the goal?

Joseph Messana: I'm not sure as a measure developer that I can answer that certainly for CMS. But personally, we know that a sizable minority of facilities have zero patients with hypocalcaemia. And so we believe that persistent hypercalcemia a problem, zero might not be an unreasonable goal ultimately as a clinician who's trying to do, you know, as many of you are trying to do the best that you can.

Whether the measure has - whether the incentives that have been provided by having this measure in the portfolio have resulted in the maximum amount of improvement or not, I don't know or you think you're asking are we getting to a threshold where, where you can expect much better. I don't know what the answer is to that question.

I expect that it'll be something that the Standing Committee would take up if this measure makes it through the methods group.

Jeffrey Geppert: Yes. And it's not a burdensome measure. It's based on clients, right? So…

Joseph Messana: It's based on CROWNWeb data collection which is part about 75% or 80% of that is (unintelligible) submitted by the labs associated with large dialysis organizations. It's part of the kind of the monthly or at minimum quarterly panel of labs that are all submitted to CROWNWeb.

Jeffrey Geppert: Right. Is it, in theory possible to target the data collection more specifically? You know, is it so it's, you know, 500 facilities that have poor performance is it the same 500 facilities, year after year?

Joseph Messana: There is some overlap independently, I think it would depend upon your threshold and the values have been changing, the values have been declining significantly since the measure was implemented so that answer were probably changed. We've only gotten 2015 I believe was the first year it was publicly reported.

So we got a small number of relatively small sample and the values have been declining over each of the years.

Jeffrey Geppert: Thank you.

Poonam Bal: People please mute your phone where it is sort of ringing in the background. It might be us. Sorry let's just try to continue talking, thank you.

(Andrew): Well, do we have any additional discussion to reliability of this measure -- 1454?

Paul Kurlansky: For reliability? No.

(Andrew): Excuse me? We're trying to get this fixed. We're calling the conference company. Well, if we don't have any more discussions on reliability the - we can go ahead and vote on Number 1454.

Paul Kurlansky: I had a question on risk adjustment in this particular…

(Andrew): Yes, you could bring that up. I should note that this one didn't pass with a high moderate rating. So we won't be re-voting on validity unless you request that we do but go ahead if you want to.

Paul Kurlansky: Well, I'm just curious is there - might there be - the reason why it's sort of a theoretical thing I don't know a lot about the physiology here. But I was just curious as to whether or not there might be any evidence that would suggest the correlation between mortality and elevated calcium was stronger in men versus women and blacks versus whites. If, you know, in other words, the assumption is that the relationship is the same. I'm just not sure that that's the case. But it might be.

Joseph Messana: This is (Joe) Messana, we do not have a specific evidence that addresses that the associations were done at the facility level and the mortality manager if used is highly risk adjusted and standardized mortality ratio, which takes into consideration individual unmeasured facility characteristics by stratification, cause of ESRD, complicated age splines, race etcetera.

And so it would be difficult to tease it out based on the facility level. So I can't answer that question.

Paul Kurlansky: Okay.

(Andrew): Okay, any…

Paul Kurlansky: We're voting on reliability here?

(Andrew): Yes. We will be voting on reliability unless there's any additional comment to be made. We can vote on reliability for our number 1454. And that will I think, finish us for the set of, you know, measures and I'll turn it over to Michael Abrams for one additional measure that we're going to talk little bit about here.

Michael Abrams: Okay, thanks (Andrew) so this is Michael Abrams here at NQF. So we're going to power through the ring here, try to stay focused shifting from kidney to brain and lung here with one measure that we actually have put on a do-not-discuss list. But then we decided to bring it up for your attention in this call. And that measure is 3492. Emergency Department use due to opioid overdose it's at the end for those of you on the subcommittee it's the very last measure on your form.

On Page 23 if you have the PDF in front of you so again it's of measure 3492 Emergency Department use due to opioid overdose. What I'm going to do is give you an overview of that measure. It's been a while since I've seen it, and then focus you on some specific questions. And I will also notify you about why we have held it for the end here.

But before I begin, can I ask…

Paul Kurlansky: Can I make a recommendation maybe that we can hang up and redo the call?

Michael Abrams: Can we do that?

(Andrew): If others are willing to do that. That might be the only fix at this point.

(Karen): Who is that?

Poonam Bal: Hi there. We actually have an operator who's is something to take care of this. Is there any way you guys can give few minutes?

(Karen): Sure. Let's just give it two more minutes and then…

Man: Oh yes, operator did it.

Operator: And this is the Operator.

Man: It's very good.

Operator: Yes I think I muted that line. I didn't want to interrupt the call I'm going to unmute some of the lines that I muted here so that you'll be able to resume your conference call.

(Andrew): Thank you.

Operator: You're welcome.

John Bott: Can you stick around for a bit and make sure it doesn't happen again.

Operator: All you have to do is Star 0. And then an operator will join and mute that. But I think I muted the line right now, it should be okay. I will not leave the conference. You're welcome.

Michael Abrams: Thank you very much.

Operator: You’re welcome.

Michael Abrams: All right, good. So again, Michael Abrams here at NQF. So we’re discussing now the final measure for this subgroup on Page 23 of your PDF, it’s 3492 emergency department use due to opioid overdose. And I think we have the - we might have some of the developers on the line from Yale. Is anybody on there? Can you just identify yourself briefly?

(Elona Richmond): This is (Elona Richmond), I’m from (I-CORE at Yale).

Michael Abrams: Okay, great. Thank you (Elona) for being on the call. And we’re going to go through this measure. I will describe why we held it for the end as I go through it. But it’s been a while since you’ve heard about it as I said. So I want to give you a little bit of an overview and then focus you on the questions that remain from your review you at, it was consensus not reaching on both reliability and validity and I’ll discuss that in turn.

So this is a new measure, description is it’s a claim space measure that captures the rate of emergency department visits for opioid overdose events using diagnostic codes, ICD diagnostic codes. Events are measured per 1000 person years among Medicare beneficiaries over the age of 18 residing in geographic areas of interest which are the unit of analysis either state or county.

It’s an outcomes measure as I said, claims-based Medicare Fee-for-Service data was the source principally for the testing that was done from the year 2009 to 2012 and they also use for validity purposes, the national emergency department sample in order to look for overdose events and that external source.

As I said, we're talking about community-level analysis, regional state and county principally state across five different states, Maryland, New York, New Jersey, Pennsylvania, Virginia and then county level data in Maryland. Before I described the results from reliability and validity, let me just skip down to some information about what they demonstrated in terms of meaningful differences.

Maryland, so in their meaningful differences analysis, Maryland does look above average compared to the five other states they had. And here's a number for you to think about in terms of that. If I've interpreted the results properly 2.16 overdose events per 10,000 person years covered is the Maryland rate that's on Page 15 of the testing form, if you want to look for that.

In New Jersey and Pennsylvania, as example competitors, it was about 1.5. So, you know, 25% less rate. Similar issues or variability was observed when they looked at say, Baltimore City, which has 5.3 per 10,000 person years versus say, Carroll County and more rural area of Maryland 0.36 so they demonstrated meaningful differences there.

So now let's, there were no exclusions reported. The other thing I should point out before we discuss reliability and validity is that they didn't do risk adjustment for the measure per se because or when the measure was implemented and tested because empirical testing and secondary source citations, even though they demonstrated social factors were influential, there was an upgrade ability and uncertainty about that they felt.

And they argued that the measure was applicable irrespective of those so they didn't actually do any social risk factor adjustments that they instead argue with evidence that this measure should be applied equally across different socio demographic populations.

Now, moving back up on page, the middle Page 23, the ratings for reliability. You all had had not reached consensus one person rated at high, too moderate and too low. And let me just give you some details about that. The reliability that they did use the atoms are score across five states and found and they also reported ranges as well and found that on the state level the items was 0.86 or higher across the various years and state.

And so they were comfortable with those, those results suggesting good ability for the measure to discriminate between the state level of analysis with regard to the county level in Maryland. There are 25 jurisdictions that we're talking about. The atoms are always exceeded 0.7. I will point out however that it dipped is low at least in the early years as low as to 0.41.

And so it wasn't quite as strong for at the county level data but the developers still argue that these were reasonable levels and of course, even in the county, they exceeded sufficient levels on an average of 0.71 being the average reliability there.

So let me pause there and see if anybody on the committee has any questions or comments about the reliability presentation of this measure.

Paul Kurlansky: I had a question about the specification. The numerator is based on events, but it's very unclear. Suppose one person has multiple events, are they talking about events or, you know, the events per population is not able to distinguish between multiple events for a few individuals and multiple events spread out over many individuals.

Michael Abrams: Yes, good. Good question. Could the developer clarify that for us, please?

(Elona Richmond): Yes, absolutely. Yes, the measure does not distinguish between multiple events and a single individual or singles and multiple individuals. And we actually felt like that was appropriate because an overdose event is an overdose event whether it happens the same person multiple times or to multiple people.

We'd want to capture that name anyway because each overdose event is an undesirable outcome and associated with downstream risk of morbidity, mortality costs, et cetera. So, and each overdose ended is in some ways an opportunity to intervene. So we felt like it was reasonable and appropriate to capture multiple events.

Michael Abrams: Very good. Any other comments or questions from the committee about reliability?

John Bott: This is John Bott while we're on specifications. I thought the denominator definition was confusing. In the MIF form, it first talks about basically all Medicare Fee-for-Service cases in the denominator. But then it goes on to start to define an emergency department visits. So it's not clear if the denominator is, you know, account of everybody in a space who's Medicare-Fee-for-Service or Medicare Fee-for-Service cases that resulted in a ED visits.

Then on the numerator side, the Excel file that was attached, there's really some disconnect going on. There's a table I think, the table specifically the tab, ICD 9, ICD 10 crosswalk codes. Seems like it's attempting to define numerator events for opioid overdose but some of those codes clearly are not in opioid overdose, one is our 09-0.2 is a respiratory arrest.

So I think there's a lack of some description going on. You know, I think you have to have an opioid overdose code and it could be supported by a code like that. But it doesn't really clearly define the numerator in that Excel file so those were a couple of concerns with both numerator and denominator.

Michael Abrams: Yes, good point, this is Michael here at NQF. So let's break it down, the second point I think we'll, let's hold for validity. But regarding the first point because the developer offer us some clarity there?

(Elona Richmond): Yes, absolutely. So the denominator is person time of eligible Medicare beneficiaries, which essentially everybody with Part A, age 18 and over. The numerator includes all overdose events that results in emergency department visit.

So in claims data, there's not a sort of simple or straightforward way of capturing all emergency department visits because how emergency department visits our bill depends in part on whether the person was discharged in the emergency department or admitted.

And then even within that there's some complexity around the kind of care they received in the emergency room. For example, it can be billed as critical care or emergency services and so on. So actually, some folks at core had developed an algorithm for identifying emergency department visits using claims data and there are other algorithms out there.

But essentially, we, first and the numerator identify visits in the emergency department from inpatient and outpatient claims. And then from among those visits, we identify those that are associated with a diagnostic code indicative of opioid overdose.

Michael Abrams: Can I - so this is Michael at NQF. Let me just clarify a point if I may that came up.

(Elona Richmond):: Yes.

Michael Abrams: So it’s plausible somebody could overdose and the substance is unclear, or that they could overdose and something besides an opioid. And they might have a history of opioid addiction and/or abuse in another code are those potential threats to validity in terms of the identifying specific opioid admission?

(Elona Richmond): Yes. So this is a central challenge and measuring opioid overdose using claims. So we could either opt for a very specific but potentially incentive definition or a much broader definition that may be more sensitive but less specific. And there’s been actually some work in the published literature around which combination of codes offers the best sensitivity versus specificity.

As you might imagine, so there’s a core set of ICD nine codes that indicate things like poisoning from heroin, poisoning from methadone. Those codes are very specific to opioid overdose and are almost never used for events that on, you know, chart review do not occur the opioid overdose.

But as you point out, there are lots of cases where people come in with other conditions or unconscious. And the reason for their condition is not immediately apparent or could be due to multiple substances or a combination of events. So in consultation with CMMI for whom we’re developing this measure, we opted for a measure that we think balances sensitivity and specificity.

So this gets to the second part of that previous question, how is the measure outcome defined? There are two ways that the outcome can be defined. One is, if an emergency department is associated with one of those very specific codes, poisoning for methadone, poisoning from heroin.

The second is a broader set of codes that are probably more sensitive but potentially less specific. So in that second definition, we include cases where a patient might come in with, for example, respiratory depression, plus a history of opioid use disorder.

I will say that those codes contributed only sort of minimally to the number of cases that we identified, the most of the cases and the measure are from the more specific definition.

Michael Abrams: So let me stop you there. So is there in your application, did you quantify for us, I don’t remember. If you quantify for us, how much of a threat to specificity there is, just a yes or no would be fine. And then we can look forward if it’s there and if it’s not there, we know it’s not.

(Elona Richmond): We didn’t, although we reference some of the literature that answers that question.

Michael Abrams: Very good, John, I think you posed that question. Does that address both of your questions about numerator/denominator specifications?

John Bott: I would have to read it now in a different way and think about it. But just in general, it’s not. It’s not a very clearly-defined numerator and denominator and having read hundreds and hundreds of technical specifications, which I think leads to the risk of different people picking this up and unevenly applying the measure specification and that's the spirit of why measure specifications is in the reliability section.

((Crosstalk))

Jeffrey Goppert: This is (Jeff) I have a question. I have…

Michael Abrams: Go ahead please.

Jeffrey Goppert: …a question. So in the denominator of your statement it says the denominator, the population at risk for opioid overdose and then the definition of examiners bed for everybody. So is that true I mean so is that so some of the variation across counties could be attributable to the health systems in those counties. But some of it could be attributable to variation and people that are actually at risk.

So question is did you actually consider more refined denominator of people that are, that have a nonzero probability of an opioid overdose?

(Elona Richmond): Yes, I mean, right on some level. It's a deep philosophical question who's actually at risk for opioid overdose? We did consider a more narrow definition or excluding certain populations at least and again, in consultation with CMMI, we opted for the broadest possible definition.

Jeffrey Goppert: Okay.

Paul Kurlansky: I had a, it's just sort of a minor technical questions. This is all based on ICD 9 data. Do we have any evidence that conversion to ICD 10 is would result and accrue in the same outcomes in sensitivity, validity et cetera and sensitivity, specificity.

Michael Abrams: Paul, thank you for bringing that up. That's actually the reason that staff hold the measure because they did not do testing in ICD 10 and so we are still trying to decide at maintenance whether or not we are going to carry the measure forward but we appreciate you considering that and since you brought up that question because the developer briefly tell us why they didn't do ICD 10 testing at this point even though they have specifications for ICD 10 identification of this measure.

John Bott: Okay and to add on to that so in the testing form of the comment that we didn't anticipate the reliability and validity to change over time but then in the actual results there is a very clear trend in reliability. So it seems contradictory. So if I could press that too.

(Elona Richmond): So beginning with ICD 9, we chose to develop a measure in ICD 9 because there is really two reasons. One is there has been some other work by other organizations developing ways to measure opioid overdose using ICD 9 codes and probably the most had been done in ICD 9 so that's where we started, and secondly we wanted to look at that over time and particularly to know that overdose rates are increasing over time. And so we wanted to make sure that general trends that we know exist every time. The newspaper that is reflected in our data and it is with Maryland.

It's true that reliability increased over time, I think again that reflects of that rate, over dose rates are increasing over time which reliability, so sorry for that contradictory statement but you are correct.

Michael Abrams: How far back can you go -- now Michael and NQF -- and get ICD 10 now? If it's 2016 that's good right.

(Elona Richmond): Right so looking ahead, we use as we mentioned the National Emergency Department Sample and a separate measure for the definition of opioid overdose developed by ARC in a distinct population for validation, and that data does exist probably for 2016 so one approach to validating in ICD 10 would be to use that 2016 data from ARC.

Michael Abrams: Okay any other – thank you for that. Any other questions about reliability here now before we take a re-vote on that if you would like to do that, the question is about reliability.

(Karen): So this is (Karen). I just wanted to point out and Michael alluded to this, I think I want to make it a little bit more clear, we had originally pulled this measure and we were actually going to not discuss it and ask the developer to bring it back next cycle, because they did not do testing with patients and data. We've had for the past year and a half roughly maybe a little bit more, we have indicated that as of 2019, we would want to see testing all of ICD 10 based measures, with ICD 10 based data.

So I think if you think about this measure and re-voting and think about this requirement that we've had in terms of testing and whether or not you will feel comfortable kind of replacing what the testing from the ICD 9 rather than the test, hopefully that makes sense.

(Elona Richmond): And just you jump in for one moment, and if possible you would be happy to submit the data from 2016 using that I first alluded to the data for the next consideration, I mean if this measure were to go ahead.

Michael Abrams: Yeah we have two cycles for a purpose, so we want to just stick to that, we do have mini measure beside yours so just want to clarify that but we understand that this is the NQF. So for the committee it is just about reliability, questions about reliability specifically.

John Bott: So the reliability testing does not propend on the ARC data, is that correct?

(Elona Richmond): That's right.

John Bott: Yeah so you could do 2016 ICD 10 reliability test without using ARC. I wouldn't use the availability the updated data the current state of your reliability testing.

Michael Abrams: Any other comments about reliability or questions? Okay hearing none, so why don't you take – let's do a re-vote on reliability for this measure, so I'll remind you the measure number 3492, emergency department used to do the opioid overdose.

Paul Kurlansky: We are only voting on reliability?

Michael Abrams: Yeah and we will talk about it in a moment and then vote on that separately. But I am going to go ahead, so yes could you do vote on reliability for 3492 and while you are casting your ballot there I am going to move on to validity and just remind you about that, and see if there are any questions or comments or questions we've already and start to talk about that in terms of the numerator definition.

So one issue was specifically overdose, so we talked about that, but let me just remind you what kind of validity testing they did, they used the NEDS, this National Sample of Emergency Department events and compared it against the Medicare claims in order to see if they have got similar numbers and really what they were doing in my view seemed to be was numerator comparisons, and these were data between 2009 and 2012 just to remind you, and numerator events in a given year something in the order of 817 for a given year, 817 in 2009 and they increased to 1632 in 2012, that's based on the measure.

You can compare that to the NEDS data that was pretty close, 850 and 1750 events respectively so very close correspondence between the NEDS and the Medicare claims suggesting that there was an external source or kind of goal standard with regards to the overdoses that they were detecting, they reported a correlation coefficient, was reported presumably with those large numbers significantly. Again this was ICD 9 data testing that was done.

No county level breaks for that kind of looks, it was done all exclusive reporting and as I said they didn't do any risk adjustments, but they justified why they didn't. Any comments then or additional points of clarification regarding validity testing on this measure from anybody on the panel.

Paul Kurlansky: All right sort of the fact that I mean NEDS as I understand is sort of an all payer database whereas Medicare is Medicare population.

Michael Abrams: They did narrow their pull to Medicare am I correct, can the developer clarify that?

(Elona Richmond): Yeah we actually compared rates in NEDS and rates in our population and absolute events in the Medicare component of NEDS and in our population and in both cases we got pretty similar results. The definition was reassuring even with the kind of different formulation of opioid overdose and kind of the approximately the same population, we get very close to the same numbers.

Paul Kurlansky: Okay so that is to sub-specify the Medicare population.

(Elona Richmond): Yeah.

Paul Kurlansky: No I just wondered because opioid problem is a little bit more concentrated in younger patients so I wouldn't necessarily expect the correlation between NEDS and Medicare but…

Michael Abrams: Yeah go ahead.

Jeffrey Goppert: I am not sure exactly what's being tested here so I am sure I understand NEDS would be considered in kind of the goal standard, they are both based on administrative data but basically same claims data just one set of claims data goes to the Medicare and another set of claims data goes to the state and that is the data from the ARC. So I don't see them as being different.

I mean the algorithm, so the claim that some of the algorithm is considered as goal standard and you are somehow validating your algorithm for identifying these -- and is that – it seems like the validity issue here it is – are these ED visits for opioid or not, yeah like….

Paul Kurlansky: I had the same problem, I had exactly the same problem.

Michael Abrams: Yeah the developer, can you maybe comment?

(Elona Richmond): Yeah that's a fair point, so there are two things, one is that there was a study by from UCSF Christopher Rowe and Phil Coffin that looked at various accommodations of those using the review and essentially the kind of core of the measures had a very high specificity, step number one, two yeah I mean it is a fair point you are comparing one claims with this measure against another, but one that was externally validated, we have nothing to do with how the data are collected, how emergency departments are identified and so on and so our measure we have to make decisions about each of those things, and we still sort of arrive at a similar conclusions or similar results within that measure.

But fundamentally is the ARC measure getting at opioid overdose, it is in many ways still a kind of measure, but one that little bit at least developed independently from ours.

Jeffrey Goppert: Yeah I mean that's something right like you said I mean there is two groups follow an independent process and coming to a similar conclusion but some validation that the process was valid. Another thing that you might have done, there is no real explicit kind of quality idea here, it's like why are you other than just the fact that some counties might have higher risk population than others, and you've kind of discounted that by just including everybody in the denominator, there is no statement about why you would expect one area, and this is for CMMI I guess it will be one I don't know whatever ACO or whatever, whatever entities.

So that's one question, and you defined this at a county level, I assume that this is going to be applied at some sort of group level and that whether anything that you have done at the county level justifies any kind of an inference about application to like a medical group or something is like a totally different question.

But there is normally a statement there why you would expect variation in this population rates to be different and if you have sort of an hypothesis to test it to say characteristics of County A and characteristics of County B that are related to the efficiency of the healthcare system or whatever it is you are trying to get at with this measure, and then show that those differences in quality actually impact the results that's kind of what we would expect I think from the validity test, of the metric.

(Elona Richmond): Yeah it's complicated because there is a whole kind of econometrics literature dedicated to which policy is improved at the population of opioid overdose and you could imagine it's a really hard question because there is a lot of – and as you said the population of highest risk have the highest opioid overdose, so it's a challenging sort of question to tease out but what we do know that there are lots of things from the health sampling that can be done to reduce overdose for most which include providing medication treatment, to some extent review things the availability of opioids and then treating opioid overdose it is a separate question and it's not actually really covered by this measure like how to reduce mortality from overdose.

But it's actually a very complicated question like how to prove the policy ideas and prove opioid overdose, but what we do think is important to be able to measure to identify the areas the communities that are have greatest needs and detractions overtime which is the sort of the goal of the measure…

((Crosstalk))

Paul Kurlansky: So this sort of goes to the rational for the measure, which was left blank and sort of left me a little bit confused, which is suggesting, in other words if this metric is meant to be applied in order to identify a problem that's one thing because this metric is meant to be used as a quality metric for the care and community then that's just completely different thing.

Because in the first case you wouldn't necessarily want to risk adjust because you are just identifying a problem and then you can know where to look and try to understand it, and on the other hand if you are using it as a quality metric then you have a huge problem not risk adjustment because of one county you are – some population, whatever population group you are looking at as individuals who are at higher risk, then they are going to look worse for even though the quality of their services may actually be much more sophisticated and better than another county where the problem doesn't even exist.

Michael Abrams: All right this is Michael here from NQF, this is a good discourse and certainly related to the validity of the connection between the measure and quality so the developer should take note that when and if you bring it to the standing committee and in your evidence presentation you might get clear what that link is of course. Are there from the committee we are on the set time here so I want to be efficient in wrapping this up. Any other specific questions about validity testing on this particular measure here such as it is presented, at this point any other questions or comments before we go to a re-vote on that.

Bijan Borah: This is Bijan, so I have again it is more like, as I understand you used data from 2009 to 2012 and then you talked about that there is a bridge between the ICD 9 and ICD 10 I guess why are you even talking about that in the existing form because I think that is the data that you know until 2012 will be pre-ICD 10 data right in NEDS?

(Elona Richmond): Yes all of the 2009 to '12 is ICD 9 data.

Bijan Borah: That is documented and maybe that is whatever and it is confusing, I am trying to understand why you are even talking about it and that if you are using the data tool from 2012 that would have any ICD 10 diagnosis to begin with.

(Elona Richmond): It may have just been a broader statement about the measures, certifications but yeah all of the testing had been done in ICD 9.

Bijan Borah: Okay.

Michael Abrams: Okay any other questions about validity before we take a vote here and close up the call. All right, good hearing no objections then, let's have then do a re-vote on measure 3492 and this is on validity, measure 3492 on validity and while you are doing that I want to thank you very much for your attention and to the developers as well who joined very last minute for us, and I will hand it back to Poonam to close out the call.

Poonam Bal: Okay. So this is the end of the call, we are done with all the sub-groups, once we get your voting results in, we will know what measures are moving forward. So thank you everyone for your time, and we will know the full results from your evaluation as soon as we can.

Any questions before we let you go.

Bijan Borah: Not for me. We are done.

Man: Thank you.

Poonam Bal: Perfect thank you so much for all your hard work.

Man: Thank you.

(Karen): Thank you all, bye.

Operator: Thank you, please standby.

END