National Quality Forum
Scientific Methods Panel Spring 2022 Cycle Measure Evaluation Web Meeting
Wednesday, March 23, 2022

The Panel met via Video Teleconference at 1:00 p.m. EDT, David Nerenz and Christie Teigland, Co-Chairs, presiding.

Present:

David Nerenz, PhD, Henry Ford Health System; Co-Chair
Christie Teigland, PhD, Avalere Health; Co-Chair
John Bott, MBA, MSSW, Consumer Reports
Daniel Deutscher, PT, PhD, Maccabi Healthcare Services
Marybeth Farquhar, PhD, MSN, RN, American Urological Association
Jeffrey Geppert, EdM, JD, Battelle Memorial Institute
Laurent Glance, MD, University of Rochester School of Medicine and Dentistry
Joseph Kunisch, PhD, RN-BC, CPHQ, Harris Health
Paul Kurlansky, MD, Columbia University, College of Physicians and Surgeons; Columbia HeartSource
Zhenqiu Lin, PhD, Yale-New Haven Hospital
Jack Needleman, PhD, University of California Los Angeles
Eugene Nuccio, PhD, University of Colorado, Anschutz Medical Campus
Sean O'Brien, PhD, Duke University Medical Center
Jennifer Perloff, PhD, Institute of Healthcare Systems, Brandeis University
Patrick Romano, MD, MPH, FACP, FAAP, University of California Davis
Sam Simon, PhD, Mathematica Policy Research
Alex Sox-Harris, PhD, MS, Stanford University
Ronald Walters, MD, MBA, MHA, MS, University of Texas MD Anderson Cancer Center
Eric Weinhandl, PhD, MS, Fresenius Medical Care North America
Susan White, PhD, RHIA, CHDA, The Ohio State University Wexner Medical Center

NQF Staff:

Matthew Pickering, PharmD, Senior Director
Hannah Ingber, MPH, Manager
Gabrielle Kyle-Lion, MPH, Analyst

Also Present:

Kathleen Balestracci, PhD, MSW, Yale CORE
Valery Danilack, MPH, PhD, Yale CORE
Elliott Main, MD, California Maternal Quality Care Collaborative
Stephen Schmaltz, PhD, MPH, The Joint Commission
Rebecca Smith-Bindman, MD, University of California San Francisco
Lisa Suter, MD, Yale CORE
Christine Walas, MSN, RN, The Joint Commission

Contents

Proceedings

(1:00 p.m.)

Welcome and Recap of Day 1

Dr. Pickering: So we'll get started, it's 1:00 o'clock p.m., on the Eastern side here. Again, my name is Matt Pickering, and welcome back, everyone, for the second part of our Measure Evaluation proceedings with our scientific methods panel.

So we're going to pick up off where we left off yesterday, and there's one more measure we will need to vote on and there's a couple measures that are pulled for discussion as well.

So we do have a tight agenda, so we're going to try to reserve each of the measure discussions to about 30 minutes, just to kind of keep us on time.

But we'll go to the next slide. Yep, and just a reminder of the housekeeping items, I won't spend too much time on these slides, but if you have any technical difficulties with the platform or logging into Poll Everywhere, for example, please feel free to directly chat the NQF staff in the WebEx platform chat box, or you can email the project box at methodspanel@qualityforum.org.

And there's of course some dial in information as well is provided with the WebEx link that you can use. The raise hand feature, we definitely utilize that and we'll recognize you as we see you pop up, and of course the chat box as well.

Going to the next slide. Just some ground rules, again, no rank in the room, we ask that you remain actively engaged and participate in the meeting discussion today.

Be prepared, having reviewed the measures beforehand to engage in the discussions.

And the base evaluation and recommendations on the measure evaluation criteria and guidance, for being familiar with those criteria and our guidance for evaluation of these measures.

We want to keep the comments concise and focused, especially for today. So if there's anything new you'd like to add that's already been discussed, we welcome that. But if it's something that is not new, it's already been discussed, we kindly ask to just refrain from providing any comments there just so that we can continue to move through the agenda.

And be respectful to others as well as share in your experiences and this is an opportunity as well to learn from others, as we definitely provide recommendations to our developers also.

Next slide. Okay. So for the agenda for today, we'll do a small recap of day one after we go through some introductions -- not introductions, excuse me, roll call, rather -- and then we'll go into the Measure Methodology Discussion.

And so, part of that is going to be 0716e and the measure developer for that is the Joint Commission, so we'll start with that measure, then we'll go into the two measures that are up for discussion.

And lastly, we'll have public comments as well and then adjourn with some next steps.

Next slide. All right, so I'm just going to do roll call, so we did Disclosures of Interest yesterday, for those folks that weren't in attendance yesterday, but will be in attendance today, we just ask that you kindly state your name, your organization, and if you have anything you'd like to disclose.

And so, for the others that were in attendance yesterday and went through this, you can just say you're present as we go through the list of names here.

All right, so starting with our co-chairs. Dave Nerenz?

Co-chair Nerenz: I'm here.

Dr. Pickering: All right. And Christie Teigland?

Co-chair Teigland: Yep, I'm here.

Dr. Pickering: Okay. Matt Austin?

Okay. John Bott?

Member Bott: Yep, here.

Dr. Pickering: Thanks, John. Daniel Deutscher?

Member Deutscher: Hello, I'm here.

Dr. Pickering: Great. Okay, Marybeth Farquhar?

Member Farquhar: Yes, I'm here. I'm from the AUA and I have nothing to disclose.

Dr. Pickering: Thank you, Marybeth. Jeffrey Geppert?

Member Geppert: Present.

Dr. Pickering: Thank you, Jeff. Larry Glance?

Member Glance: Present, thank you.

Dr. Pickering: Thank you. Joe Hyder?

Joe Hyder?

Okay. Sherrie Kaplan?

Sherrie Kaplan?

Okay. Joseph Kunisch?

Member Kunisch: Present.

Dr. Pickering: Thank you. Paul -- Kurlansky, excuse me?

Member Kurlansky: Not too hard, I'm present, thank you.

Dr. Pickering: Thanks Paul, I know, I always want to say Kurlnasky. Thank you, Paul. So, Zhenqui --

Member Lin: Yep.

Dr. Pickering: Zhenqui Lin?

Member Lin: I'm here.

Dr. Pickering: Thank you. Great. Jack Needleman?

Jack Needleman?

Okay, we'll circle back for Jack. Eugene Nuccio?

Member Nuccio: Here.

Dr. Pickering: Thank you. Sean O'Brien?

Member O'Brien: Here.

Dr. Pickering: Thank you. Jennifer Perloff?

Member Perloff: Here.

Dr. Pickering: Great, thank you. Patrick Romano?

Member Romano: Present.

Dr. Pickering: Thank you. Sam Simon?

Member Simon: I am here.

Dr. Pickering: Thank you very much. Alex Sox-Harris?

I think Alex is going to be running a little bit late today. Just one more time, Alex Sox-Harris?

Okay. Ron Walters?

Member Walters: Present.

Dr. Pickering: Thank you. Terri Warholak?

Think she probably wasn't be in attendance today. Terri Warholak?

Eric Weinhandl?

Member Weinhandl: Present.

Dr. Pickering: Thank you. And lastly, Susan White?

Member White: I'm here.

Dr. Pickering: Thank you so much.

Member White: Thanks.

Dr. Pickering: All right. Okay, so, and we'll circle back on some of the others that didn't say that they were here when we start getting into the measure discussions.

Okay, and it looks like we have quorum for our measure that's up for vote, which is 0716e, so thank you all.

Okay, we'll go to the next slide. And I'll turn it over to Hannah to do a recap of Day One. Hannah?

Ms. Ingber: Thanks, Matt. Yeah, just a quick recap of what happened yesterday, we reviewed seven measures, one of them got pushed to today. And I'll just go over the results real quickly.

There were only two measures that did not pass on either, reliability or validity, one measure did not pass on reliability but was CNR on validity, one measure passed both, reliability and validity, and two measures passed reliability but were CNR on validity. There was one measure that was CNR on both, reliability and validity.

And we'll move forward with the discussion of 0716e today, which got shifted to this afternoon. Thanks, everyone.

Measure Evaluation

Subgroup 1 - Renal

Dr. Pickering: Great. Thanks, Hannah. Okay, we'll keep on going, so next slide, please.

All right. So, before we get started on this last measure for evaluation, or for voting on here today from yesterday, I want to just double check to see if the Joint Commission is on the call. Do we have the Joint Commission on?

Ms. Walas: Yes, this is Chris Walas from the Joint Commission, we are here.

0716e ePC-06 Unexpected Newborn Complications in Term Newborns (The Joint Commission)

Dr. Pickering: Great. Thanks, Chris. Okay, so this is still a Subgroup 1 measure, you can see that reliability was a No Pass and validity was a Consensus Not Reached, so we will be looking at both of those today.

Our lead discussants are Paul, and also Sam as a secondary, and again, the measure developer is the Joint Commission, you can find this on page 9 of the discussion guide.

So, similar process, I'll present the measure and then I'll note both, reliability and validity testing. I'll turn it over to Christie to have the lead discussants then present any concerns related to reliability first, discussion around reliability, and then the developer provides any comments to the questions related to reliability, then we vote on reliability.

And then we'll go to validity where the lead discussants will then present any concerns for validity, we'll go through that same process, okay? All right.

So, this measure is Measure Number 0716e ePC-06, it's Unexpected Newborn Complications in Term Newborns, this is a New Measure but it's an eMeasure of a chart-based version of the measure, and this measure is a hospital level performance score reported as the rate per 1,000 full term newborns with no preexisting conditions who had Unexpected Newborn Complications, typically calculated per year.

It is an outcome measure, so it uses Electronic Health Data, it is at the facility level of accountability. For the reliability, it did not pass, with a low, insufficient rating.

The developer conducted reliability testing at the encounter level. An encounter level of validity testing served as the reliability testing for this measure. So the results and methods are noted under validity testing, so that's where I'll go to next.

And so, for your assessment of validity, keep in mind you'll be evaluating the validity of the data element level, the testing there, for your assessment of reliability.

So, for validity, during the validity testing, 61 sample cases were successfully re-abstracted from 14 hospitals. And then four hospitals from the original pilot were not included in the validity sample due to multiple hospitals having the same accreditation identifier.

During the virtual visits, site staff shared their screen, navigated through the EHRs of the sampled patients, while The Joint Commission staff manually re-abstracted each data

cesarean Re-abstraction findings were compared with the original electronic data submission and any disagreements were adjudicated with reasons for discrepancies noted.

So for the testing methodology, it included all clinical data elements and all

editable demographic elements were scored. All measure data were re-abstracted with original data having been blinded so that re-abstraction was not biased. And re-abstracted data were compared with original data on a data element by data element basis as well as by measure result. So measure agreement and data element rates were calculated, and clinical and demographic data were scored separately. The measure agreement rate was corrected for chance variation with kappa statistic.

So 32 records across eight different hospitals in Site 1 exhibited a match rate of

93.8 percent. Twenty-nine records across six different hospitals in Site 2 exhibited 100 percent match rate in measure outcome. The overall kappa was .99 -- or .955.

There was an exception -- there were some exceptions to this agreement, the secondary diagnoses and the procedure codes were lower since they were not always collected according to instructions. And the gestational age, author date/time, and birth weight were low due to differing data sources. The demographic variables of race and ethnicity also had lower agreement rates for Site 2, which was due to

different data sources.

So, going further down, I'm just going to talk a little bit about some of the potential threats to validity.

So, missing data elements are counted as mismatch. So, for Site 2, there were no mismatches from missing data. For Site 1, three data elements accounted for most of the missing

data. The missing race and ethnicity codes for Site 2 are due to different data sources as well.

For some of the exclusion analyses, the developer compared the frequencies of the denominator and numerator by site before and after the exclusions. And no formal statistical

test was performed for the effect of exclusion on the performance score.

Denominator exclusions ranged from 4.8 to 56.7 percent, indicating variability throughout the sites.

The developer stated that risk adjustment was not needed due to the three exclusions, babies with congenital malformations and genetic diseases, babies with pre-existing fetal conditions such as IUGR, and babies who were exposed to maternal drug use in-utero.

So to further enforce, or reinforce, the need to not risk adjust this measure, the developer presented conceptual evidence. And the developer stated that to guard against potential

overcoding and undercoding, babies with a length of stay greater than 5 days will count as a moderate complication even if they do not have complication codes. So, in addition, the developer stated that risk adjustment is not included in the maternal conditions because this would add burden to collecting the measure. And then lastly, some maternal conditions are complications of labor that affect the baby, which is what the measure is trying to assess.

Okay. So with that, I'll turn it to Christie to see we have a discussion on reliability and the concerns related to reliability testing, and then we'll vote on that, then we'll move to validity. So Christie, I'll turn it to you and our lead discussants.

Co-chair Teigland: Yeah, thank you, Matt. Yeah, these are such important measures obviously because of where we rank in the United States on these issues, so I think these are, you know, clearly important measures.

We did talk about the cesarean birth measure yesterday and the issues with reliability and validity that the SMP, you know, really felt were pretty fatal to the measure.

We're going to now talk about the unexpected newborn complications, and I'm going to turn this over to Paul first to talk about any -- some of the issues I know are similar but there may be some additional issues, or different issues with this measure, so, Paul, if you could start with the discussion of issues around reliability, that would be great.

Member Kurlansky: Great. Thank you very much, Christie, and thank you, Matt. It was a great summary, much better than what I would've done.

But, in any event, so this is a measure of birth complications among full term singleton deliveries among babies without congenital malformation, genetic disease, or preexisting conditions such as intrauterine growth restriction or babies who were exposed to maternal drug use in utero.

And there's a hierarchy of diagnoses that would include the baby in the numerator without double counting, which means that, you know, if there was a more serious complication then the fact they have a less serious complication does not double count them.

The data relies on the Electronic Medical Record and it makes use of SNOMED logic identifiers, names, and codes, and ICD-10, and perhaps other sources as well.

It's noted, there was no specific reliability testing was performed but instead there was the validity of the data element. Data element validity score was performed using chart extraction with data from 61 cases in 14 hospitals that used two different EMR systems, one was Epic and one was Cerner.

Kappa was excellent for the score level testing and potentially good for the data elements testing.

So, potential issues, there are a few related specific -- more specifically actually to how it impacts reliability and then, I guess we'll delve into those and then subsequent to the vote we'll go into specific concerns regarding validity itself.

First of all, it's not so much a, maybe a problem, but a perspective and that is to realize that testing here is with sites that are -- testing is voluntary. And therefore you have to assume that this is the -- the figures achieved were, have to be the most optimistic estimate of what would actually happen if this measure were put into effect and it became compulsory upon sites.

So it's just something I think -- there's no way to quantify that, but it's something just to keep in the back of your mind.

Now, there were 61 cases from 14 hospitals, and we don't know the exact breakdown, but that gives you about four cases per hospital. So we have an extremely limited ability to be able to assess either, intrahospital or interhospital variability, there just is not enough -- we don't know the breakdown, and even if we did, there's just not enough to be able to say anything. So, from a reliability point of view, it's a little bit concerning.

Now, I stated that the 61 cases were quote unquote, statistically representative of the total sample of 6,699 cases, however I couldn't find anywhere where the statistical method that was used to arrive at this representative sample was disclosed. Nor do we have any evidence that the sample was in fact statistically representative of the entire population, it may well have been, it's just there's no evidence one way or another.

Many of the reviewers point that out that not all elements were tested, and in the initial application, none of the elements in the denominator were tested, which is a huge potential issue because, as Matt pointed out in his presentation, it's between five and 56 percent of patients actually had exclusions due to the denominator.

However, in the response from the sponsors, there was testing of many more elements and, including denominator elements, which showed similarly good agreement.

I leave it open as a question for the developers, and I'm very glad that they're here, to just inform us, was that subsequent testing with the same sample as the original testing? Because, if so, I don't know why it wasn't included in the original application.

Or was there a subsequent retesting, was this one site that went back and was retested, where did that data come from? And you may have told me, I just missed it, but I was just concerned about that.

And, you know, this is a -- there's another concern that one of the reviewers brought up, which I thought was actually very interesting, and it sort of relates to all of these electronic metrics, and that is that the data comes from several places, it comes from SNOMED, it comes from LOINC, it comes from ICD-10, and maybe from other things.

And this may be just due to my ignorance regarding the internal workings of the quality data model, but there may be differences depending upon where the data is coming from within the Electronic Medical Record as to how reproducible it is.

In other words, the data from SNOMED may not be as reproducible as something which is a little more quantitative, such as LOINC or ICD-10, and there was no differentiation in the presentation as to the source of the data and how it may or may not have impacted the correspondence between what was reported and what they found in the chart.

And then, you know, lastly, an issue which was not so much an issue here, but was a potential issue raised yesterday, you know, Epic is 35 percent, Cerner is 25 percent of the market. So, you know, you've got about 40 percent of the market that's not represented here.

And this is, I guess it's also more of a general question regarding these e-metrics, are developers obligated to test this in multiple different Electronic Medical Record systems to make sure that all of the

Electronic Medical Record systems, or at least the majority of those four major which would (audio interference) compromise about 75 percent of the country.

Is that -- this is more of a question for us, is (audio interference) method panel, is this sort of a requirement that we should think about, you know, given that there may be a variability in the ability of different medical records, (audio interference) Electronic Medical Record systems, to be able to respond to the needs of the different metric.

So those are the, I think the major issues regarding reliability that I think were raised by some of the reviewers as well as my own concerns. Sam, what did I miss?

Member Simon: I don't (audio interference) --

Member Kurlansky: We lost you.

Co-chair Teigland: Yeah. Sam, you can unmute --

Member Simon: (Audio interference) mentioned --

Co-chair Teigland: There you go. (Audio interference) in and out.

Member Simon: Oh, sorry about that. No, Paul, you covered the waterfront and then some, that's pretty much what I saw in the comments, including some of my own.

In terms of your question about EHR coverage, I believe that the NQF standard is two, at least two different EMR systems, so in this case the developer did meet the standard. But anyway, I don't have anything to add, Paul.

Co-chair Teigland: Okay. So, let me, I think I saw a hand up. Larry, do you still have your hand up?

Member Glance: I do. I have a very quick comment to make, I think that it is very difficult to establish validity, data level validity, and enhance, in this case, data level reliability, when you're looking at 60 or 70, or 100 charts.

And, in particular, because the incidents of severe newborn morbidity is under five percent, you really can't validate the outcome, I don't think that you can validate the outcome of interest. At least, I'd be interested in hearing what the measure developer has to say about this.

But in an outcome measure that is not risk-adjusted, granted there are exclusions, but it's not risk-adjusted, I think it's particularly, and in any outcome measure actually, it's particularly important to be able to validate the most important data element, and that would be the outcome.

Co-chair Teigland: Okay. Thanks, Larry. I don't see any other hands up, so maybe I'll turn this over to Chris, you know, comment about the additional testing that was done, why wasn't it included in the original submission -- one of Paul's questions.

And then, if you could talk a little bit about how the source of the data might impact the reliability testing that you did do.

I think maybe the issue about, you know, are they obligated to test every single, you know, EMR, we can maybe put off to a later conversation, but let's talk about those two things first, Chris, if you could.

Ms. Walas: Thank you. Yes, this is Chris from the Joint Commission. So I will touch on a few of those topics and then I will turn it over to my colleagues to address some of the others.

So, first, as far as the missing data elements, those data elements were all tested at the same time during the pilot testing. When we pulled the pilot testing results into the document, we just didn't have, unfortunately all of them represented.

But we, you know, were easily able to go into our files and pull the testing results to provide them to you, so we appreciate the opportunity to do that. But they were all tested and, you know, run through the EHR at the same time as the original data elements were presented.

So, all data elements were tested, we did highlight what data elements were used for demographics, denominator, denominator exclusions, or numerator. Generally the agreement rates were excellent on all of them.

We also broke down, you know, we tend to just provide high level data, but we realized the severe and moderate complication rates would be important to you so we did provide the breakdown as the sites also get of the severe and moderate complication rates, to distinguish between the two.

So, even though, you know, this was a small sample we do feel that the measure outcome and the kappa in the validity study did indicate that the numerator cases were being captured correctly.

We do understand, these are rare conditions, you know, so with the testing that we were able to do, we do feel that they have excellent agreement rates and that they were being captured correctly to be used in the measure.

I will turn it over now to Stephen to talk about the sampling methodology, and some of the more statistical conversation, so this is our statistician, Stephen Schmaltz.

Mr. Schmaltz: Good afternoon, the way the cases were sampled is, there was a stratified random sample. We wanted to look at an equal, around an equal number of cases per hospital, so the cases were spread over the hospitals.

And we also wanted to evenly divide the numerator and the denominator cases, so we basically over-sampled the numerator cases. So we had about the same numerator cases as we looked at denominator cases, and spread evenly over all the hospitals in the study.

Co-chair Teigland: Do you have any thoughts about how the source of data -- did you look at differences between the sources of data and reliability?

Mr. Schmaltz: Well, at least the data that I looked at mostly used ICD-10 codes, they did not tend to use SNOMED codes, so I can't really --

Co-chair Teigland: Okay.

Mr. Schmaltz: Comment on the difference between SNOMED and ICD-10.

Co-chair Teigland: Yeah.

Ms. Walas: This is Chris again. So, the SNOMED codes, we do use a tool that selects the SNOMED codes based on the ICD-10 codes. So they are applicable to the matching ICD-10 codes which we are testing.

As far as the measure not being risk-adjusted, you know, we do find that this measure is correlated highly to the chart based, which you did mention that this is an electronic version of the chart based, and so we do follow a lot of the same principles and we work closely with our technical expert collaborator, Dr. Elliott Main, and he's here and can speak a little bit more on how the risk-adjustment impacts -- or, not risk-adjusting impacts this measure. Dr. Main?

Dr. Main: I didn't know if you all wanted me to go now or in the second section for the discussion --

Co-chair Teigland: Yeah, I was going to say, if we could defer that to the validity conversation --

(Simultaneous speaking.)

Ms. Walas: Sure.

Co-chair Teigland: That would keep things a little cleaner. So, any other comments on reliability from the SMP, either, you know, the subgroup that evaluate the measure or others -- see any other hands up.

Member Geppert: I just have a very basic question. So, of the 61 cases included in the study, how many had the measured outcome?

Member Simon: Thirty of them.

Member Geppert: Thirty of them?

Mr. Schmaltz: At least the original ones, they had the outcome, and they were in the numerator.

Member Geppert: Thank you.

Co-chair Teigland: Any other questions?

Member Glance: Quick comment, what was the kappa statistic for the numerator?

Co-chair Teigland: I'm not sure they -- did they provide kappa?

Mr. Schmaltz: We didn't do it separately for the numerator and denominator. What we did was, with kappa, for those where the re-abstractor said, they were in either, the numerator and the denominator, which was most of them.

And, Chris, do you have the kappa that we had for that?

Member Kurlansky: The kappa for the outcome was actually extremely high.

Co-chair Teigland: Right.

Ms. Walas: The total kappa was .955.

Co-chair Teigland: Yep, very high.

Member Kurlansky: And 914 for one site and 100 for the other site was reported.

Ms. Walas: Correct.

Co-chair Teigland: Last call for anymore comments on reliability, or questions?

Dr. Pickering: I see Patrick Romano's hand's raised, Christie.

Co-chair Teigland: Oh, okay, Patrick?

Member Romano: Yes, good morning. I'm in the other workgroup here but I think this just raises some interesting questions that we may need to discuss separately.

You know, where we all know, those of us who are in this business, that doing this testing of eCQM measures is difficult and it really requires an unusually close collaboration between the sites and the measure developer, and in this case the Joint Commission has described, you know, how they had to do that.

It is a burdensome process, it is a resource intensive process, and yet it often seems like we end up with barely adequate data to say anything. You know, with small number of sites and so, it just leaves us in a quandary about how to approach the problem, so I think it's just an issue that we're going to need to discuss a little bit more to figure out, you know, what are our expectations.

Co-chair Teigland: Yeah.

Member Kurlansky: It's a very important point, and it relates to the first point that I brought up, and that is that the sites that the data emerges from are voluntary sites, and it will be -- if this gets accepted and becomes mandatory on sites, will there be that level of performance, is it reasonable to expect that level of performance from all sites, and I don't know the answer to the question.

You're right, though, it's more of a general question, I think, than a specific one for this particular metric.

Co-chair Teigland: More data definitely would be better.

Member Kurlansky: More data -- but yes, in this particular case, more data would be much better. Because, you know, 61 cases, they were selected for, you know, for a high portion of outcome with interest.

But still it's only 61 cases and, you know, about four cases per hospital, so the interhospital variability, intrahospital variability is really not something that can be tested here.

Member Romano: And ideally we'd like to see confidence intervals surrounding these estimates of agreement, sensitivity, specificity, kappa, whatever you want to use. But ideally, we'd like to know, you know, are we pretty confident that those confidence intervals don't include, you know, ridiculously small values.

Member Kurlansky: And, you know, and if you bid it for the 61 it might even seem reasonable, but if you bid it for the four, the four, four, and four, you know, it might be much less reasonable over the 14 hospitals.

Co-chair Teigland: Probably most definitely the case, which is why no within hospital variance was reported.

Member Kurlansky: Yes.

Co-chair Teigland: Anything else before we move to a vote on reliability?

I guess we will do that. Gabby?

Ms. Kyle-Lion: All right, everyone, I will go ahead and pull up my screen. Just a reminder that this is only for Subgroup 1 voting and that Joe Kunisch is recused from this measure but he is not in this Subgroup so that should not affect your voting.

All right. Voting is now open for Measure 0716e, or 0761e, sorry, on reliability. Your options are A for moderate, B for low, or C for insufficient, and I believe we are looking for nine votes here in the denominator.

(Pause.)

Ms. Kyle-Lion: I am seeing eight votes. I'll just give it another second to see if we get that ninth one.

Okay. I am still only seeing eight votes, but that's okay because that is quorum for this measure. I am going to go ahead and close the poll now.

Okay, voting is now closed for Measure 0716e on reliability. There was one vote for moderate, three votes for low, and four votes for insufficient, therefore, the measure does not pass reliability.

I will pass it back to Christie and Matt.

Co-chair Teigland: Okay. So no change in our vote for reliability. I guess we will move on to validity. Paul, do you have additional comments about validity?

Member Kurlansky: Yes. I think basically two comments. One is that the measurement tested for validity internally in terms of comparing with the "gold standard" of chart extraction.

However, I didn't see any evidence that it has been tested for external validity, i.e. is this metric actually measuring quality.

And, you know, there is no -- It's a new measure, so as I understand it the requirement would only be for face validity, but I didn't see any formal testing for face validity.

That then goes into the second concern, which started to seep into conversation inevitably anyway, which is the absence of risk adjustment.

I would point out that this measure is distinctly different in this regard than the companion measure regarding cesarean section, because the cesarean section the outcome is, we had this discussion yesterday, the outcome of interest is actually a decision.

Here the outcome of interest is not a decision. It is a medical fact of what happens to, of what happened to the baby.

So, you know, and clearly I find it, you know, I mean I understand cardiac surgeons are not the first people that you would want to ask about neonatal health, but I find it very difficult to believe that there are not maternal conditions other than those excluded in the denominator that might impact the outcome of interest here, which is quite a large series of claims.

It's not only the most serious, but it's also a whole, you know, hierarchy of outcomes, and I really find it very difficult to believe that even amongst those full-term, you know, women, otherwise healthy women that there are not conditions which might impact this outcome.

So to me the absence of risk adjustment is a -- And the rationale, you know, in the cesarean section was actually apparently detailed rationale if you followed it as to basically demonstrating that a lot of the difference that is seen amongst sites is actually due to surgical decision rather than, in that case, in that analysis was BMI and age.

Here there was no similar sort of analysis that was presented and basically it was that it would be too burdensome for people to collect this information.

So I mean if it's too burdensome to collect the information it's too burdensome to have the metric in the first place, so it did not resonate with me.

So those are the two issues, and they are related, that concerned me regarding validity. Sam, I --

Co-chair Teigland: Sam, yes, anything to add?

Member Simon: Nothing to add for me. I thought that was great.

Co-chair Teigland: I don't see any other hands. Do you, Matt?

Dr. Pickering: Yes, Larry --

Member Simon: Larry's.

Dr. Pickering: -- has his hand raised.

Member Glance: Yes. So I just want to second Paul's comments. There are a fairly extensive number of maternal conditions that would be expected to be important risk factors for newborn morbidity, things again like maternal BMI, whether or not the mother has had a required cesarean delivery, any placental abnormalities, whether or not that mother has had a prior C-section and then co-morbidities.

So I do think, and I am sure Dr. Main will answer this, but I do think a priori that risk adjustment is appropriate and I think that the lack of risk adjustment is likely an important threat to validity. Thanks.

Co-chair Teigland: Anyone else before we go back to -- Dr. Main, maybe you want to comment on the risk adjustment now since that's a big issue here.

Dr. Main: Sure. And thank you for allowing me to join. I was not really part of the e-measure development here, but I am the measure developer for the parent measure which has been endorsed by NQF for the last three years without risk adjustment based on ICD-10 codes.

So this is in play already in America and used extensively in California probably for about six, seven years, over about four or five million patients.

This is something we report on the hospital level to every hospital in California and the Joint Commission is doing this nationally.

So here's the challenge as to what you want in an e-version transition from a current version of the measure. This, contrary to what Larry may say, there is most of the effects on the infant at term are going to be based on morbidities that are brought prenatally rather -- and those are the ones we strove to exclude except in -- Well, a lot was to focus on what happens on labor and delivery.

Now one of the challenges here is that this is then a measure of two people on two different patients, the mother and the baby. It is incredibly hard for most hospitals to link those together let alone report those externally.

I work with Epic extensively and it's only been in the last year or two that there is any kind of linkage between the two patients within their medical records, but to extract those is hard.

Nonetheless, in California we were able to do that because we have extensive link datasets for many years and we did do an analysis looking at age, BMI, let's see, race, education, insurance, parity, prenatal care, prior cesarean section, maternal hypertension, maternal diabetes, and in -- Let me just show you what we got when we did that.

We did this in two different ways. One was predicted over observer, over expected, and the other is, and that's in orange, and in versus observed or expected in green dots, compared to the "X" which was the observed.

So this caterpillar plot shows the distribution, again, of 220 hospitals. This covers about 450,000 patients, of which 82 or 83 percent are the denominator for this measure.

And you can see that there isn't much change when you do adjust with, you know, as -- And look at the orange and green dots here. They are pretty similar to where we are with the "X" except as you get into the higher rates.

There you start seeing some differences as you get above 2, 2-1/2 percent. There is kind of inflection point here around 2, 2.1 percent where the curves do differentiate.

But this made us feel very comfortable that for most of the distribution here that the numbers are really, exactly correlative. We don't believe that there is a big difference, you know, between 3 and 4 or 5 percent that there is -- We looked at this as a high rate versus a normal rate or an expected rate here.

We don't compare hospitals, you know, your -- You know, a three is better than, or a two is better than a one. Excuse me, a one is better than a two, et cetera.

But we really want to be able to follow hospitals along and use this primarily to show that as you change the cesarean birth you are not changing the baby outcomes.

So by and large actually this is used internally within hospitals to compare their time over, their progress over time.

So given that you really can't combine mother and baby records, two different patients, which I don't think this is an issue with any other measure that you have in NQF where you are looking at two different patients together in one measure.

So given --

Dr. Pickering: Dr. Main, apologies to interrupt. Matt from NQF. I was wondering if we could try to keep close to our time here if possible to -- Try to do a little bit more of a summary with this.

I will just mention as well, you know, we talked about this yesterday with risk adjustment with the SMP. It's definitely in the purview to evaluate threats to validity, and that includes risk adjustment.

Some of the decision making on whether risk adjustment is appropriate for certain outcome measures, there may be considerations that the Standing Committee should evaluate that, you know, the "importance" of it.

Those concerns raised today from the SMP that are more clinically focused, especially if risk adjustment is or is not needed, would be those concerns we would share with the Standing Committee to weigh in on.

So I just wanted to circle back on that because we talked about it yesterday as well and just wanted to mention that. And sorry to cut you off, Dr. Main, but I was just trying to see if we could sort of get through, sort of to move to a vote, sorry.

Dr. Main: Okay.

Co-chair Teigland: No, that was really useful data though, Dr. Main. Thank you for sharing that.

I guess I would just ask for maybe a comment on the other key point that Paul made from whoever on the Joint Commission, and that is about no external validity testing was done.

What was the rationale or was there and we didn't see it? Thoughts on that? Chris?

Ms. Walas: Sure. Stephen, do you want to talk a little bit about any additional statistical analysis that you did?

Mr. Schmaltz: Yes. So we actually did look -- Yes, I can talk to that. We actually did look at the correlation for these hospitals of their eCQM measure versus the other PC measures that we collect on both eCQM, well, not eCQM, but the chart-based measures that we send to the Joint Commission as part of their accreditation requirements.

But keep in mind that this is a low percentage measure and less than a year worth of data whereas the chart-based measures were based on a full year's worth of data.

When we did the correlations we did not find any significant correlations with the other PC chart-based measures except for PC-06, but I think mainly we just need more data to kind of look at that correlation again.

Co-chair Teigland: Yes, okay. Thank you. Any other comments?

I am not seeing hands raised, but I'm not always seeing them. All right, I think that being said let's move to the vote for validity. Gabby?

Ms. Kyle-Lion: All right. I did just want to make one correction. We did receive one additional vote via chat for reliability. So for 0716e the reliability results were two votes for moderate, three votes for low, and four votes for insufficient.

So it still is no pass, but we did want to make everyone aware that we received and additional vote.

With that being said I will move into the validity vote. Okay. Voting is now open for Measure 0716e on validity. Your options are A for moderate, B for low, or C for insufficient, and, again, we are looking for nine votes here.

(Pause.)

Ms. Kyle-Lion: Okay, we are at nine so I will go ahead and close the vote. Voting is now closed on Measure 0716e.

There were two votes for moderate, six votes for low, and one vote for insufficient, therefore, the measure does not pass on validity. I will pass it back to you, Christie and Matt.

Dr. Pickering: Okay, great. I also want to thank you very much for including the voting from yesterday's measures and thank you again to the Joint Commission.

Again, apologies, Dr. Main, for interrupting as we were trying to continue moving through the rest of our agenda today.

Measure Methodology Discussion

Subgroup 1 - Perinatal and Women's Health

Dr. Pickering: But with that I think we are going to go now to our next series of measures. So we have two measures that were pulled for discussion.

So, Gabby, if we could put up the first measure.

Ms. Kyle-Lion: Okay.

Dr. Pickering: And, Jack Needleman, you are on the line, correct? Can you hear me okay?

Member Needleman: Yes, I am.

Dr. Pickering: Okay. So both of these measures that we will be discussing today did pass with pass with preliminary reviews from the subgroups.

The first one here is Subgroup 1. So Jack Needleman will provide a summary of any concerns related to I believe it was risk adjustment for this measure.

I want to just mention that, again, for the risk adjustment components if there is anything that is more clinically focused or a decision of including certain factors in the model, we definitely want to document those concerns and share them with the Standing Committee for their consideration.

So keeping that in mind -- well Jack will discuss any concerns he has with the measure and then we'll open it up for any further discussion from the subgroup members and see if there is any recommendations or additional concerns we can note as well for the Standing Committee.

3687e ePC-07 Severe Obstetric Complications (The Joint Commission) Patient Safety

Dr. Pickering: So you can see that the measure listed here is 3687e. It is the Severe Obstetric Complications measure. It's also a Joint Commission measure and it's located in the discussion guide on Page 14.

So, Jack, I will turn it over to you and maybe just sort start out with the concerns for this measure that you have --

Member Needleman: Sure, Matt.

Dr. Pickering: -- and then we'll turn it back to Christie to facilitate any discussion with the subgroup.

Member Needleman: Okay. Thanks, Matt. And given the time we have I will try to be briefer than I usually am.

First of all, severe obstetric complication is a critically important issue in the U.S., much higher than there should be levels of maternal mortality and other morbidity.

Nothing I am about to say underscores, is meant to take away from the importance of the measure.

I do have some questions about the measure, some of which can be deferred to the Standing Committee, but also some real technical issues about whether the measure developers, whether the factors, some of the variables that are included are simply right and should be excluded.

So with respect to the broader issues I would kick to the Standing Committee, 80 percent of the -- Well, the measure documentation lists a couple, one or two dozen, I've got to admit I don't have the count in front of me, complications.

When you look at the distribution of the complications 80 percent of the complications are transfusions and that was not highlighted at all in the documentation, but it is something the Standing Committee ought to think about because it's heavily -- Much of the variation we are seeing, even in the unadjusted data, is around transfusion rates.

The model does include social risk factors in the risk adjustment model. I will also leave it to the Standing Committee to decide whether that is appropriate given the nature of this.

But there is one variable in particular in the risk adjustment model that I think is, it shouldn't be there, or I am concerned shouldn't be there, and seems to having an outsize effect on the risk adjustment, and that's the measure of economic housing and stability.

This is an ICD-10 measure that was first introduced into the ICD-10 coding in 2016. It was substantially revised in 2021. I cannot figure out which version was used in the testing or development of the measure from the documentation we received, but it's rare.

In the cases in which this was done the rate was only one-tenth of a percent. I have looked at some of the housing surveys and six-tenths of the adult population says that they are at risk of either foreclosure or being evicted and 4 percent say they are not, very likely they are not going to be able to make their next mortgage payment or rent payment.

So given that, this number just looks low, and the Medicaid percentage in the population suggest that it may be low, which means it may be an unreliable measure to include in the risk adjustment model right now given the way hospitals are collecting it. That is one concern.

There are only 62 cases that are here and when you look at the risk adjustment modeling there is some reason to believe that this model, this variable is contributing to over-fitting of the risk adjustment models even though it has been, even though it may not be a reliable measure.

They present two risk adjustment models, one for all the cases and one for the 20 percent of the cases that are not transfusion cases.

The OR and the not transfusion model is five, which is an extraordinarily high odds ratio, much higher than we see in most cases, and when you look at the performance of that risk adjustment model we see a range of cases that go from zero to 81, the predicted or the expected after risk adjustment go from 50 to 51, except for one case which is predicted at 55.

I have never seen that kind of compression in a risk adjustment model and it just screams at me it's over-fitting.

Now you don't see the same compression in the full model with all the transfusion cases, but there is still a fair amount of compression in that model and given how aggressively the risk adjustment model predicts the non-transfusion cases, all the variation we are seeing is the unadjusted -- I think all we're seeing is the unadjusted variation of the transfusion cases in the performance of this model.

Again, that might be deferrable to the Standing Committee, but the unreliability of this one measure which seems to be having a disproportionate effect on the risk adjustment model, five OR in the non-transfusion cases, 1.8 in the full model, just feels to me like it shouldn't be there.

This variable is unreliable and should not be in this risk adjustment model for technical reasons, not just it's a social determinant model.

I really would like to see the risk adjustment reset before I would approve this measure. That was my concern.

Co-chair Teigland: So given the compression, did they show ability to differentiate good and bad performers? It sounds like maybe that's not going to be possible with that type of a --

(Simultaneous speaking.)

Member Needleman: Well, certainly there is no variation of performance associated with the non-transfusion-related cases.

Co-chair Teigland: Okay, all right.

Member Needleman: There is some variation in the risk adjusted predictions for the full sample, including the transfusion-related cases.

And, you know, it's a sample of, you know, ten or 20-some odd hospitals and, you know, the question, that gets back to some of the earlier discussions we have had about in a select group of hospitals picked for testing how much variation do we anticipate, unexplained variation do we, or what hospital level variation do we expect to see.

So I think those statistics are there. I don't remember them, but it's the issue of over-fitting in the risk adjustment model that just grabbed me and said I am really unhappy with the technical specs here for including what I think is an unreliable variable in the risk adjustment model and the effect I am seeing it have in the risk adjustment model.

Co-chair Teigland: Any other -- Any comments on what Jack said or questions?

Let's go to the Joint Commission to respond. Tell us a little bit more about that variable. Is that Z code variable, because I do know they are not used at all --

(Simultaneous speaking.)

Ms. Walas: Sure. This is Chris from the Joint Commission. So the Z codes that we use are the Z95 codes and that is the economic and the housing instability value set. There were 62 encounters or 0.1 percent that had that code.

The American Hospital Association is encouraging for providers to use codes and did release new coding guidance that the social determinants of health can be assigned based on information documented by all clinicians involved in the care of the patient hoping to increase the amount of Z codes being used.

Z codes for homelessness were among the most used code in that Centers for Medicare and Medicaid service report that showed 1.6 --

Co-chair Teigland: 1.6 percent, right.

Ms. Walas: Yes. So homelessness was one of the top used codes. So as far as the rest of the risk model I will turn it over to our colleagues from CORE, Dr. Katie Balestracci, and then, Katie, feel free to call on whoever from your team who can best answer their questions.

Dr. Balestracci: Yes. Hi. I am going to actually invite Valery Danilack, who is part of our team here at Yale CORE who led this measure.

Dr. Danilack: Hello. This is Valery Danilack.

Co-chair Teigland: Hello.

Dr. Danilack: Hello. So to first answer the question about the over-fitting of the model, so we did see relatively equal amounts of, relatively equal area under the curve with both the model with and without transfusion only cases.

So that led us to believe that the model, the risk adjustment model for the severe obstetric complications excluding transfusion only cases was not, you know, entirely predicting that outcome with the risk variables alone.

Member Needleman: Valery, before you go on, on that issue, is the table wrong? Again, the risk adjusted models, risk adjustment values range between 49 and 55 and basically the interquartile range is 50 to 51, and you've only got one case above 51, which is 55.

That looks like a massive over-fitting to me, notwithstanding the C-statistics. Is that table wrong?

Dr. Danilack: That table is correct. The risk standardized rates are calculated from a predicted outcome which is a model that includes the risk variables with an addition for a random effect for the hospital sites divided by just a risk prediction model with the risk predictors without that hierarchical model for the hospital sites.

So it gives a sense of how much of the variation is from the risk factors versus from the -- What additional variation on top of the risk factors is from the hospital sites, and that is multiplied by the overall rate in the population.

So the starting point for the risk standardized is the average in the population and it is adjusted from there.

Member Needleman: Yes, but you've got a hospital with zero cases risk adjusted up to 50. You've got a hospital with 81 cases risk adjusted down to 51.

Dr. Danilack: So we note that the prevalence of the outcome is low. As you noted it's, you know, quite -- The prevalence of the outcome is low, and then we have data from about 25, from 25 hospitals for this testing.

So because the prevalence is very low we don't expect very wide variation in the measure scores and, you know, given that these hospitals are all Joint Commission hospitals we do expect more variation once more hospitals, both in number and in a variety of hospital characteristics, are tested.

Dr. Balestracci: If I may? This is Katie Balestracci from Yale CORE working on this measure as well.

Also just noting that because this is a commonly used way of communicating severe maternal morbidity, in this case severe obstetric complications, we are reporting or suggesting the reporting of this measure as a rate of 10,000.

So I just want to remind the Committee that 50 out of 10,000, right, when it is translated to a hospital that has 500 delivery encounters per year or 700 delivery encounters per year, is going to be a very, very small number.

So part of what is happening here is the translation into a rate, which again has been chosen because this is a common way in the field that these types of complications are discussed, is going to look like less variation than may actually be going on.

And, again, as we expect on implementation with a large number of hospitals then being included in the calculation of measure scores and the impact of the greater variation that we would see then in this outcome greater variation as well.

Co-chair Teigland: So with the data we tested with we really can't see any meaningful differences between the hospitals but we are expecting to see that with more data?

Member Needleman: Yes. On the non-transfusion complications.

Co-chair Teigland: Yes, right, right.

Member Needleman: Although there was variation in those in the RoR rates across the facilities.

Is somebody from the Joint Commission going to speak about 80 percent of the broader risk adjustment model and whether you think that's performing and specifically whether a variable that is I think unreliably reported should be included in the risk adjustment model?

Ms. Walas: This is Chris from the Joint Commission. So all of our risk adjustment variables were taken from the co-morbidity study by Leonard, et al.

So those variables have been widely tested and accepted as risk factors. And so that is where we, you know, decided on those overall risk factors.

Member Needleman: Yes, but I noticed in your documentation you did not look at, you did not specifically report testing the looking at the accuracy of the reporting of the economic housing instability measure.

Part of my concern is it's under-reported. Until hospitals do a better job of reporting it it's not ready for prime time for inclusion in a risk adjustment model even if we believe economic and housing instability is a risk factor for maternal morbidity and should be included in the ultimate risk adjustment model.

The question is whether it should be included in the model now given the inaccuracies in it. I did not see any evidence in the documentation you provided that you had tested, you had reviewed that measure.

I saw a lot of documentation for the other measures but not for that one.

(Simultaneous speaking.)

Dr. Balestracci: Valery, is that -- Oh, go ahead.

Dr. Suter: This is Lisa Suter from CORE. I will just jump in to also acknowledge that although there is recommendation from the AMA and the AHA to report this variable and we anticipate that it is currently under-reported, you know, having a variable in a measure is a very strong incentive for hospitals to report this and we know that housing instability from empiric evidence outside of this measure development work is predictive of maternal and fetal outcomes.

So while we acknowledge there is under-reporting, we feel very strongly that not including this in the measure would be unwise.

The Joint Commission and, you know, others who might use this measure have annual review processes that allow re-evaluation of the measure to, you know, look for concerning trends or unintended consequences.

But in a situation where we know there are a lot of existing disparities, and I am sure the TJC can speak to, you know, their plans for stratifying the measure, you know, I think it's important to recognize that there is likely to be important variation within subgroups that will be illuminated by the measure and we think that housing instability is an important modifying risk factor to include in the model given the other intended stratification by social determinants of health. Thank you.

Co-chair Teigland: So let me just ask one more question on that. Why would we include this variable but then stratify by some of the other important socioeconomic risk factors?

Dr. Balestracci: If I may? This is Katie Balestracci from Yale CORE. The plan, and we are looking into approaches, is to stratify by race and ethnicity.

Co-chair Teigland: Okay.

Dr. Balestracci: This has been a really important decision from the get-go. It is well established in the literature and in studies that there are significant gaps in outcome particularly by race and ethnicity.

Co-chair Teigland: Yes.

Dr. Balestracci: And we want to illuminate those for hospitals not adjust for them. So that is a particular and purposeful decision based on a social risk factor, or in this case on race specifically, that has been made because of those gaps.

Adjusting the measure in addition by potential SDOH is something we considered, and as Dr. Suter just noted, landed on this particular variable given both the evidence that exists and that was the reason for kind of a distinct decision based on these two variables.

Co-chair Teigland: Mm-hmm.

Dr. Balestracci: I hope that answers your question. I am happy to --

Co-chair Teigland: Yes. No, no, it helps for sure. So you did test other socioeconomic risk factors like income, for example, in the model?

Dr. Balestracci: We did not, but this is why. There was a very careful determination made about what variables are available in the EHR.

Co-chair Teigland: Mm-hmm.

Dr. Balestracci: And this is for measure developers across the country as we move towards eCQMs in a way that takes advantage of the great breadth of clinical data in EHR systems.

We are still hampered by SDOH that may be available in these systems and surely there are a number of organizations, including NQF, looking at how to pay more attention and bring those variables into use more widely.

Co-chair Teigland: No, I just thought since you used the Z code for homelessness you might want to try the Z code for income.

There is lots -- I mean there are several Z codes I could see being impactful here, which is why I was just wondering why you just tested that one.

I know they are all completely under-reported, significantly under-reported given there are only less than 2 percent.

Dr. Balestracci: Mm-hmm.

Co-chair Teigland: But I think we only have a few minutes. Jack, I don't know if your questions were answered.

Member Needleman: I see Larry's hand up.

Co-chair Teigland: Okay.

Member Needleman: And from my perspective I have had an acknowledgment it's an unreliable variable and then the question of whether an unreliable variable should be included in a risk adjustment model is one that this Committee could answer or we could defer it to the Standing Committee.

But the issue is very clear, it is an unreliable variable, it is under-reported. It seems to be having a disproportionate effect in the risk adjustment models.

The question is, you know, should it be there in an endorsed measure.

Co-chair Teigland: Yes. Larry?

Member Glance: Yes. I just want to make one more quick comment, and I think this is not going to be for discussion today, but I think it's something that we ought to consider at some later time.

I think as a Committee, as a Panel, we have spent a lot of time thinking about whether or not to risk adjust for socioeconomic variables or whether to risk stratify.

I would just like to make the point that if you stratify, so for example if you were to separately report this particular outcome measure for black mothers as opposed to white mothers the issue that you might have is that the overall percentage of black individuals in the population is about 11 percent, so your outcome, your denominator for your black mothers would be about a tenth of the population that you have overall.

And since many, many of our measures are based on hierarchical modeling and shrinkage estimators it is very, very likely that when, if we did stratify any measure on race that because of the shrinkage that you would end up losing a lot of the variability that you currently see in terms of variability across and between different facilities.

It's something that -- Again, I don't think this is the time or the place to really discuss this because we are running out of time, but I just wanted to bring that up since Jack was focusing on this one particular data field which was meant to take into account differences in social vulnerability between patients.

Co-chair Teigland: Right. Right. All right. Barring any other thoughts, comments, I think we will, you know, just note this as an unreliability issue that we will, you know, include in our comments to the Standing Committee and we will leave this as pass/pass. Jack, are you good with that?

Member Needleman: As long as the issues are clearly articulated to the Standing Committee I am happy with it.

I appreciate the efforts from the developers to have an honest conversation about the rationale for including things and the limitations to the measures that were there.

Co-chair Teigland: Right. Let's work with the SMP to make sure that documentation includes, you know, very clearly states the position. Jack, if you could help with that?

Member Needleman: Sure.

Co-chair Teigland: All right. Matt, let's move on to the last one.

Dr. Pickering: Okay. Thank you all for the discussion. Again thank you to our developer and the Joint Commission for answering the questions, and Yale CORE as well, for answering any questions from the Standing Committee.

2830 Pediatric Computed Tomography (CT) Radiation Dose (UCSF)

Dr. Pickering: We'll move to the last measure that we'll be discussing today. And thanks, Gabby, for pulling that up. So this measure is 2830. It's the Pediatric Computed Tomography Radiation Dose measure.

So this measure, the measure developer is University of California-San Francisco. I just wanted to see if the -- USCF's are you on the call?

Dr. Smith-Bindman: Yes. Hi, this is Rebecca Smith-Bindman. I am here.

Member Needleman: Great. Thank you for joining us. So I won't go too much into detail about the background of the measure since it also did pass both validity and reliability.

Alex Sox-Harris, he is on the call and he is going to be our lead discussant to talk about concerns related to this measure and then we'll open it up to other Subgroup member discussions and then turn it to the developer to respond.

Again, noting that if there is issues that we can have the Standing Committee resolve we definitely will document that and articulate that to those Standing Committee members.

So, Alex, I will turn it to you and then we can have Christie facilitate the discussion.

Member Sox-Harris: That's great. Thanks so much. Thanks to everyone for indulging my concerns at the end of a long couple day meeting.

So I wanted to discuss this measure which passed on reliability and validity for two reasons. One, I have serious concerns about the validity of this specific measure based on the way it's scored.

But also I think it's an example of how a measure can have good methods and results for reliability and empirical validity testing but still have major validity problems in my opinion.

So I am talking about validity in the most universal way. I mean does the measure as it purports to do distinguish between sites that have poor or excessive radiation from those that don't.

So the way this measure works is that there is a reference distribution of radiation dose and there are actually different referenced distributions per anatomic area and age strata and overall, but that detail is not relevant to my concern.

It is completely outside my expertise to judge the appropriateness of the reference distributions or where in the reference distributions the line gets crossed from reasonable to excessive radiation, completely leave those details to the Standing Committee.

Or yet another thing that is outside my expertise is whether the 75th percentile in the referenced distribution is the right marker for scoring. So that's all for the Standing Committee.

My concern is the way the measure is scored and how sites get designated as for excessive radiation versus acceptable.

For simplicity sake, sites whose median radiation dose is greater than the 75th percentile of the referenced distribution are considered poor or excessive.

So all sites with between 51 and 100 percent of their scans above the 75th percentile of the referenced distribution are considered poor.

To me that actually seems reasonable well enough. I think if sites have, you know, over 50 percent of their scans above the 75th percentile of the referenced distribution that seems like there is probably room for improvement.

But my concern is the converse. So sites can have between 0 and 49 percent of their scans below the 75th percentile and be considered acceptable. That's quite a large range meeting that criterion.

So imagine a site with two radiology units. In the first unit within the site it does 60 percent of the total site's scans and all of them are below the 75th percentile of the referenced distribution, okay, so automatically it's going to be considered acceptable.

Then the other unit, which does 40 percent of the site's total scans, and all of them can be above the 95th percentile of the referenced distribution, so an alarming density of high dose scans but still this site would be considered acceptable by the scoring metric.

And this, I mean this isn't my clinical area at all, but I can imagine this happening due to a poorly calibrated machine or poorly trained staff or what have you, so this seems to be a problem to me.

So, you know, as a consumer of quality data I should be able to assume that those sites considered acceptable have reasonable radiation doses, but up to 49 percent of their scans could be very, very high in the referenced distribution.

In fact, it's mathematically possible that the mean, not the median, radiation dose in acceptable sites could be higher than those categorized as poor and to me this seems like a very serious problem with the measure's validity.

So in their response to this concern the developers acknowledged this problem and defended the decision of this dichotomized scoring system was needed to achieve good reliability.

I was glad to hear they explored other more granular scoring systems, but they discarded them because they found they were not reliable enough.

So this is referencing reliability in my opinion, which is good to care about reliability over validity.

And taking the necessity of a dichotomized outcome at face value, which, you know, is debatable, I could imagine other ways of scoring this measure that might alleviate the concern.

And, you know, this is just off the top of my head, but you could consider a site acceptable if at least 50 percent of the scans were below the 75th percentile, which is the current way, the current -- But then add and 20 percent of the scans are, less than 20 percent of the scans are above the 90th percentile for something like that that protects against the concern I have against the high scans being all packed at the top of the distribution.

I am almost done here. So, again, I appreciate your forbearance. So I was trying to think of other approaches that might be explored to incorporate uncertainty into the scoring.

Currently it's just, it's the mean, so there is no confidence or no confidence interval or anything like that incorporated.

So currently a poor site with a median at the 76th percentile is judged to be fundamentally different from an acceptable site with a median at the 74th percentile.

So you can imagine a system, and we deal with lots of measures like this that incorporate uncertainty and use statistical difference from some reference point to categorize good and bad sites, which, you know, take into, have the charm of taking a variation and also a sample size.

So we are not here to discuss alternative scoring systems, but I just wanted to highlight that the limitations of the current approach could be handled by some alternative strategies.

So in summary, I will stop, is just I think the way the measure is scored has serious risks of categorizing sites that are poor as acceptable and that worries me as a quality measure nerd and as a patient and that's why I rated this measure as low on validity.

So I just wanted for the record, at least for, you know, this group and also for the Standing Committee to register these concerns. So I will pass it back to whoever is leading the discussion right now. Thank you.

Co-chair Teigland: Yes. I don't see any hands with any other comments so I am going to, Rebecca, give you an opportunity to respond and then, you know, I think we'll just probably make sure this is documented so that the Standing Committee can think about this a little bit more since it seems we do think, you know, based on what we were presented with this measure is reliable and valid based on the data you presented and maybe there is other ways to construct it, but comments, please, Rebecca.

Dr. Smith-Bindman: Thank you for the comments and thank you for the opportunity to respond to them.

I want to start by just reminding you that we score the measure in two ways and I think one of the ways that we score the measure at least does give a sense of some of the gradation in performance that I think you are concerned that we don't appreciatively highlight in this dichotomy.

So the two ways we score the measure is the overall proportion of exams that are above the benchmarks of 75th percentile, and that means for every reporting entity, a hospital.

It's the total number of exams that are above the threshold and that's a continuous score between 0 and 100 percent. We use that to judge sites that are using excessive doses if they have more than twice the expected number of high dose exams.

So if they have more than 50 percent they are excessive, but there is a continuous measurement that you can use to judge a site's performance.

So in the example that you gave that site would have 40 percent of exams that were above the 75th percentile. I understand they are very, very high above the 75th percentile, they are above the 95th, but they are still giving you a gradation that can help you understand the facility that has 0 percent versus your example of 40 percent.

So we do have that one way of scoring which I think does give the continuous score. And then the second way is we look at within individual patient age group and stratum whether or not their doses are too high by the dichotomy that you pointed out.

I want to make one other point before I get to your primary point of that we don't do a good job for sites that pass.

So the second point I just want to make is that our measure is not intended to find those gross outliers and (audio interference) about calibration.

I didn't mention this when I responded, but facilities calibrate the machines every single day. The technologists calibrate the machines sometimes multiple times a day if the machine has been off for a period of time.

The medical physicists are responsible for calibrating machine using phantoms many times a year, and so that calibration happens on a regular basis.

And then as I noted in my response, every State in the U.S. has a radiation control agency that is responsible for regulating these machines and facilities have to report doses that are in that 95 percent category.

So for the gross outliers I agree with you that they are important, but our measure measuring doses once a year is not the way to pick up those lethal doses.

So I don't think our measure is trying to get at that. We are trying to get at routine performance. You are absolutely right that if you pass our measure doesn't do a good job of separating the barely passing from the doing really super well.

I think your suggestion of putting in another caveat it's below the 50th, but there also are a more than certain number of really high exams is sort of interesting.

The problem is just one of sample sizes and that most facilities in the U.S. don't do that many exams in children and when you stretch that to the level of evaluating just a couple of potential patients our measure really loses its reliability.

We do not judge every single scan as being correct or incorrect. We judge group scans. For any individual patient or one, two, or three individual patients, there can be very legitimate reasons for being really high even as high as the 50th percentile.

And so I am afraid that if we try to bring into a measure judgments that rely on one, two, three, four patients that it's really going to introduce unfair judgment and reliability.

Basically having one patient (audio interference) those greater than 95th percentile could make it seem like they have cracked (audio interference).

I think it's we're starting to exploring, but the basic question that we tried to look at can it say if the facility is excellent, good, very good, poor, or very poor. We weren't able to do that.

It just -- It relied on a few patients that reached the categories that we really would not be giving a reliable (audio interference), whereas we felt our mission was reliable.

Co-chair Teigland: Thank you. Alex, was that helpful?

Member Sox-Harris: Yes, I mean I appreciate your responses and I understand there are complexities to it and clinical realities and sample size realities that I am not as attuned to, so I appreciate the response.

I retain my concern about the basic construction of the measure. You know, median, and just the form I would not -- Other -- I would not like to see other measures with this scoring system because of this exact concern.

So I mean we'll pass it on to the Standing Committee or, you know, if anybody else has comments about it.

Co-chair Teigland: Yes.

Member Sox-Harris: But, yes, my worries remain.

Co-chair Teigland: Which we will definitely articulate to the Standing Committee and we can also, you know, discuss this a bit more in one of our upcoming SMP meetings.

I think at this point we'll let the measure stand with the pass/pass criteria. We will pass along your concern to the Standing Committee. I think that wraps it, Matt.

Dr. Pickering: Okay. Well thank you very much to the measure developer. Thank you so much. Alex as well, thank you for raising those concerns.

It's definitely something we will relay to the Standing

Committee as well for their assessment, so I appreciate you raising that, Alex, thank you.

So with that, that does conclude our evaluation and discussions of measures for this cycle. So we are going to move to public comment.

NQF Member and Public Comment

Dr. Pickering: So we'll have some opportunities for members of the public to provide any comments on the measures that have been discussed in the past couple of days, so today and even yesterday if you weren't able to attend yesterday.

So the lines can be opened. If you are on mute and you are calling in you'll have to do star six I believe or you can raise your hand, you know, through the participant list if you are on the line and you would like to provide a comment.

But now it's an opportunity for the public. So we'll give it a few minutes and if anyone from the public would like to make a comment you can do so now.

(Pause.)

Dr. Pickering: Okay, just double checking again. Now is the time for public comments on the measures that have been through the SMP this cycle.

Any members of the public, now is the opportunity to provide your comment.

(Pause.)

Dr. Pickering: Okay. I will check in one more time. So last call here for members of the public if you would like to provide any comments to the Scientific Methods Panel please do so.

(Pause.)

Dr. Pickering: Okay. So seeing no hands raised and nothing coming through chat and not hearing anyone speak up I think we can move on from public comments.

So we will go to next steps before we adjourn. So, Hannah, I will turn it to you.

Next Steps

Ms. Ingber: Thanks, Matt. So, yes, our next steps are on the next slide. Thanks. For this submission cycle the full submission deadlines are April 4th and 11th for all the measures that we discussed today and all the non-complex measures that have not gotten through SMP.

So just a reminder for the measure developers on the line that the full submission deadlines are April 4th and 11th.

Regarding these measures that were discussed, NQF staff with summarize the relevant measure information and discussions of the SMP and provide that to the various Standing Committees, as we have mentioned today.

The Standing Committees will evaluate these measures in the June or July timeframe. Late June or early July will be when those meetings are convened.

Then the CSAC, the Consensus Standards Approval Committee, the CSAC will review these measures in the November/December timeframe. These meeting dates are still being ironed out but will be announced publicly as soon as they are available.

The next Intent to Submit deadline is then on August 1st. Next slide, please. We also wanted to draw some attention to upcoming SMP meetings for advisory calls where we discuss methods not necessarily in relation to actual measure evaluation.

So the next one is on April 27th from 10:00 a.m. to 12:00 p.m. Eastern Time, after that we have one on May 24th from 12:00 to 2:00 p.m. Eastern time, and then there is one on July 14th from 12:00 to 2:00

p.m. Eastern Time.

The topics are to be determined, but we welcome all members of the public, anyone who is interested, to join us for those calls. Next slide.

As always, you can feel free to reach out to us at methodspanel@qualityforum.org with any questions, comments. As has been mentioned in the chat, we welcome your feedback and look forward to hearing from anyone who has any questions or comments.

I will hand it back to Matt first to adjourn us. Thanks, everyone.

Closing Remarks

Dr. Pickering: Yes. Thanks, Hannah. So I will provide Christie and Dave if you would like to say any closing remarks for today. So, Christie --

Member Needleman: Before -- Matt?

Dr. Pickering: Oh. Yes?

Member Needleman: This is Jack. Before we get to closing remarks can I add an item to future agendas for the Committee?

Dr. Pickering: Sure. Is this for consideration of advisory meetings or the measure evaluation meeting?

Member Needleman: The advisory meetings.

Dr. Pickering: Okay.

Member Needleman: There has been a rather -- Larry Glance started a rather robust discussion about Bayesian shrinkage which has continued in the chat.

I don't believe hierarchical, the hierarchical methods that CMS uses in producing their expected to predicted estimates inherently use Bayesian shrinkage.

But it would be good to just solicit from some of the key CMS developers the extent to which that is being incorporated into their measures and which measures.

Jeff -- Oh, God, I'm terrible with names here.

Dr. Pickering: Geppert?

Member Needleman: Yes. Thank you. Jeff -- No, not Jeff Geppert. Jeff Silber at Penn who was on the Cost and Resource Use Committee with me used to feel very strongly that Bayesian shrinkage hid poor performance of low volume providers. It also hides good performance of low volume providers.

So to the extent that we are seeing measures with Bayesian shrinkage included in the methods, and we ought to find out how many of those we have, it would be good to talk about how the Committee feels about that as an appropriate balancing of information from the overall sample and individual providers.

And that I think is a methods discussion, not a review of measures discussion.

Dr. Pickering: Yes. Thanks, Jack. Anybody --

Member Romano: I second that.

Dr. Pickering: Yes.

Member Romano: I think it's an interesting topic to discuss and it very explicitly, as Alex just mentioned, prioritizes reliability over validity.

This is, you know, this raises sort of broader questions about when reliability and validity are in tension how do we resolve those tensions.

Dr. Pickering: Thanks, Dr. Romano. I see Larry has also provided some additional comments in the chat, I would add PE ratio in CMS measures based on shrinkage estimators that it's based on hierarchical modeling.

So thanks for raising that, Jack. There is definitely a lot of topics for consideration for upcoming advisory meetings, so we'll list this on that list of topics and then be circling back with our SMP on how we prioritize those for future meetings.

But we'll add that to the list, Jack. Thank you. It sounds like we had some agreement from others as well. Great.

Okay. Well I was going to say Christie maybe could provide some closing remarks and then maybe Dave after that.

Co-chair Teigland: Sure. I think it has been a really productive couple of days. Everyone has been very, very thoughtful, did your homework, brought up some really interesting issues that we definitely will have full agendas at those upcoming meetings, Hannah, that you described.

So we won't have a lack of things to talk about. This is an evolving art/science. It is a science, but there is also some art to it as we have discovered over the past couple of days and there is some, you know, things that are not so black and white.

So I thank everyone for all their input as always and look forward to our upcoming discussions on these really interesting topics that have emerged over the last couple of days.

Dr. Pickering: Great. And, Dave?

Co-chair Nerenz: Yes. I'll just do a couple of things here. First of all my general thanks about the dedication and the energy that people put into this work.

It takes a lot of time to think through these measures, evaluate them, go through these discussions. It's not for the timid. So I thank everyone for being involved in this and doing such a great job.

I specifically want to thank everyone for the style of our discussions. This is a remarkable group in terms of the way that we treat each other with mutual respect as we share ideas and make points back and forth. I think we treat the developers with respect.

I just want to call that out in a positive way. I appreciate it. I thank you for it. The world doesn't always work that way and I am glad that this group works that way, so it's nice.

A couple things that have come up, also echoed a little bit with our meeting last Fall, getting into some interesting deep water issues of just how we make our judgments, how we do our work, and I'll just highlight a couple.

We had a couple times this cycle and a couple times last cycle where there were some really good and interesting points made about how a measure could be made better, and I agreed with those points.

The question about our process then is at what point does a flaw that we can identify that if corrected would make something better, when is that grounds for failing a measure or should we pass it through as we have it and then point out to the Standing Committee it could be made better knowing that the developers don't have to make it better.

I don't have the answer on that this afternoon, but I think we could perhaps tee that up for a future discussion. I think we have handled it well both cycles, but it's kind of a sticky thing.

That leads me directly to some further clarification of the rating scale we use because ultimately the decisions are based on this four category rating.

I was reminded yesterday in one of the Subgroup 2 measures that we re-voted, the voting ended up out of ten votes, six votes moderate, four votes low, therefore CNR.

If one person had switched from low to moderate that measure would have passed. Now it still goes forward, you know, it's not the end of the world either way, but as I do my ratings I am constantly wondering what is the boundary supposed to be between high and moderate but far more importantly what it's supposed to be between moderate and low.

When do I fall one way and when do I fall the other? Any further clarity among us that we can bring to that in the future I think is going to help us all.

And then in general we always talk about risk adjustment. A couple things have come up in the chat about, you know, what's in our purview, what's in the Standing Committee purview.

It might be nice to work through a few examples of that just for our own clarity of saying what is worth us spending time kicking around and what do we just pass on to the Standing Committee.

So we have done our work very well, very successfully, raised a number of issues. I look forward to additional discussion.

Adjourn

Dr. Pickering: Great. Thank you so much, Dave, and thank you as well, Christie, for both of your facilitation and time and leadership for these proceedings as well as those in the past.

I also thank our SMP members as well for all of the work and review of these measures and thoughtful considerations on the approaches taken and the developers as well for all of their time in submitting, and going through this evaluation process as we do know it. It can be quite intensive.

And, lastly, just thank you to the NQF team for all the back end work that they do to get all the materials sent out, to get the meetings scheduled, and just making sure we're running through this process so that you can conduct the work that supports our overall mission, so thank you all very much.

We will be following up with some email communications moving forward and we are very much looking forward to the advisory meeting that we will be having in April, a lot of topics to kind of filter through and think through, so we'll be looking forward to that.

But with that I hope you all have a great remainder of your week and have a great weekend and we will talk to you soon.

(Whereupon, the above-entitled matter went off the record at 2:49 p.m.)