National Quality Forum
Scientific Methods Panel
Tuesday, March 22, 2022

The Panel met via Videoconference, at 10:00 a.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, Maccabi Healthcare Services
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, Duke University Medical Center
Jennifer Perloff, PhD, Institute of Healthcare Systems, Brandeis University
Patrick Romano, University of California Davis
Sam Simon, PhD, Mathematica Policy Research
Alex Sox-Harris, Stanford University
Ronald Walters, MD, MBA, MHA, University of Texas MD Anderson Cancer Center
Terri Warholak, University of Arizona, College of Pharmacy
Eric Weinhandl, Fresenius Medical Care North America
Susan White, PhD, RHIA, CHDA, The Ohio State University Wexner Medical Center

NQF Staff:

Elizabeth Drye, MD, SM, Chief Scientific Officer
Tricia Elliott, DHA, MBA, CPHQ, FNAHQ, Senior

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

Also Present:

Andrea Benin, Centers for Disease Control and Prevention
Jonathan Edwards, Centers for Disease Control and Prevention
Kevin He, University of Michigan Kidney
Epidemiology and Cost Center
Kathy Lester, Kidney Quality Care Alliance
Elliott Main
Lisa McGonigal, Kidney Quality Care Alliance
Stephen Schmaltz, The Joint Commission
Vahakn Shahinian, University of Michigan Kidney Epidemiology and Cost Center
Craig Solid, Kidney Quality Care Alliance
Chris Walas, The Joint Commission

Contents

## Proceedings

(10:01 a.m.)

### Welcome, Introductions and Disclosures of Interest

Dr. Pickering: Sorry, I was talking the whole time. My apologies. I was on mute. I was saying good morning, everyone. Thank you very much for joining the Spring 2022 Measure Evaluation Meeting.

My name is Matt Pickering. I'm the senior director here. It's a pleasure to see you all once again.

I wanted to thank you all for your time for today and also tomorrow. We do have a packed agenda, as we do every cycle with you all, but also thank you for all of your time, insight and expertise leading up to these meetings.

We recognize there's a lot of material you have to go through and assess, so thank you all very much for your continued support and engagement with this effort.

We do have a full agenda today, but before we get to that, I'm just going to touch on a couple of housekeeping items. So if we can get started, we can go to the next slide.

So just some housekeeping reminders. This is Day 1, but it will be the same for Day 2. We're using Webex. So you can definitely use the Webex platform feature or you can dial in as well. We encourage you to use the Webex platform feature if you can.

And also if you're talking, please feel free to use the video feature. We use that to be more engaging with this virtual environment. So please use the video feature.

And this also has the ability to mute yourself. So if you're not speaking, please kindly put yourself on mute just to prevent any background noise.

We also encourage you to use the different types of ways you can engage with the group. So one of those is with the chat feature.

So if you are using the Webex platform, the little chat callout icon is at the bottom right of your platform.

If you click on that, it will pop up the chat box. You can chat everyone or you can chat individual members, and you can even chat any NQF staff themselves.

You can also raise your hand. So if there's an opportunity for discussion throughout today and you want to be recognized and you just want to chime in, you can raise your hand.

We'll definitely keep an eye on the raised hand feature and call you as you are -- in order as it's been received.

But also if you are wanting to just contribute without using your hand-raise feature, you can do so by just taking yourself off mute and contributing to the discussions.

As we go through, we'll definitely ensure that we have folks available and on the call today by doing a roll call.

But if you're experiencing any technical issues and you're using the platform, you can use the chat feature to chat with the NQF team directly, or you can also email the project box at methodspanel@qualityforum.org. So that is, again, if you're having any technical issues.

We'll go to the next slide. Just a few other housekeeping items here or reminders. We do have a break today. So we do have a lunch break built in later this afternoon.

So about 30 minutes. And so we'll try to keep to that lunch break depending on how the morning goes.

We may build in additional breaks if needed, just depending on how we're going through the proceedings and the agenda today. We do at least have a lunch break built in.

We also want to maintain quorum today. So out of each of the focus groups our minimum number here for quorum is eight -- for each of the subgroups, excuse me, not focus groups.

Each of the subgroups, our minimum here is eight for quorum. So we're keeping an eye on that.

So if you have to step away at any point in time throughout the day, please be sure to message the team so that we can just keep track to make sure that we're maintaining quorum today.

I mentioned the chat feature and the raised hand feature. So we encourage you to use those throughout the proceedings today, again, raising your hand to be recognized for discussions and using the chat as well to communicate any additional concerns or issues you'd like to raise with the group.

Muting and unmuting, keeping yourself on mute until, you know, you would like to participate or be called on, and then unmute yourself and contribute verbally to the discussion.

And then not to use the speakerphone. We just kindly ask to not use your speakerphone. It just causes a little bit of a feedback loop sometimes and some issues with the audio. So if you can, try to not use the speakerphone if possible.

Also, introduce yourself, especially those who are sort of calling in on the call. Just make note who you are so that we can recognize you.

We also are having this meeting recorded and transcripts will be generated, so that helps our court reporter as well identify who the person is that's speaking.

When I mentioned the technical support, if you are having any technical support issues or need technical support, please feel free to message us through the chat feature, one of the NQF staff members, or you can email us as well.

Okay. Next slide. Okay. So now we'll just go through some introductions and disclosures of interest.

But before I do, go to the next slide, and I just want to give an opportunity -- go to the next slide there, Gabby.

There we go. Thank you. I just want to give an opportunity for our two co-chairs, Christie and Dave, to provide some opening remarks and welcoming the group for today's proceedings and tomorrow.

So I'll start with Christie, and then we'll go to Dave. Christie?

Co-Chair Teigland: Thanks, Matt. Welcome, everyone. This is our spring meeting. Spring is here, and we've got a very, very tight agenda today.

As usual, the NQF staff has done just an incredible amount of work and an incredible job organizing some really complex measures that we're dealing with.

Some issues are old to us, but some are sort of new. So it promises to be a pretty challenging day.

Just, you know, hope everyone can keep really focused on the issues and the real reasons why we voted the way that we did so that we can get through our agenda, but looking forward to the day. Thanks.

Dr. Pickering: Thanks, Christie. And Dave?

Co-Chair Nerenz: Yeah. Thanks, Christie and Matt. Really not much to add after those welcome and thanks. I just echo those, you know.

It takes a lot of work, as you all know, to go into the background, read all the things we're asked to look at, think about it, make decisions, take time these two days.

So we appreciate it very much and look forward, as always, to our discussions.

Dr. Pickering: Thank you, Dave. Now, I do want to go to the next slide, if we can. And I want to allow -- give an opportunity for our new Chief Scientific Officer, Dr. Elizabeth Drye, who has recently joined us at NQF, to provide some welcoming remarks as well.

Dr. Drye comes with years of experience. Some of you know Dr. Drye very well or have seen her in these types of convenings as -- wearing the measure developer hat.

She comes from Yale-CORE most recently and has a lot of experience with measurement science and application of measures and development.

So Dr. Drye, I'll see if you want to give any welcoming remarks to the group.

Dr. Drye: Thanks, Matt. It's exciting for me to be sitting in this chair and starting a new chapter where I'm working with all of you in a different role, as Matt said.

I was at Yale-CORE for 15 years working on and developing primarily risk-adjusted outcome measures, ECQMs, other measures that we put through NQF and through this panel.

And, in fact, I was at the Kaizen -- I think it was 2017 -- where we -- working with, you know, received this approach to setting up this panel and I just want to thank you so much for your service.

I know how voluminous the work is that we do and how critical it is to assuring scientific integrity of the work that we do at NQF.

My role here is to -- I'm responsible across the development processing measure application partnership contracts as well as a couple -- and I oversee a couple other contracts, but I'm going to be thinking with the team and you about, you know, and we have time at the end of April -- April 27th there's a two-hour meeting with this committee to do some strategic thinking about how we're structured, how to make the process worthwhile. So I'm looking forward to thinking with you about that.

Also I just wanted to note, as Matt said, I've worked with a number of you before. It's really nice to see some familiar faces. Some people I haven't connected with in a while.

Today, I'm going to be -- I'm going to be listening. I'm going to try to join the entire meeting.

I may have to step away for a few minutes here and there, and I'm just really looking forward to hearing your thoughts and getting a more, you know, current sense of how the committee is working, and again, just really appreciate and respect all the work that you do and looking forward to working with you in the coming months.

Dr. Pickering: Great. Thank you, Dr. Drye.

I also just want to recognize the other members as listed on your screen here. These are other NQF staff persons that have been instrumental in the work that we do and leading up to the meeting today.

So Tricia Elliott, being our senior managing director, as well as Poonam Bal, who is another senior director here at NQF, Mike DiVecchia, who is our project -- or excuse me, not a project manager, but a director here working with the project management team, as well as Hannah Ingber, who is our manager, as well as Gabby, who is our analyst as well. So a big thanks to them.

Okay. So we'll go into introductions and disclosures of interest. So this is probably not too foreign to everyone, but we do it every cycle with you all.

So today we will be combining the introductions with the disclosures of interest. You received two disclosure of interest forms from us.

One is your annual disclosure of interest, and the other is specific to the measures that we'll be evaluating this cycle.

So in those forms we ask you for a number of questions about your professional activities. And today, we'll ask you to verbally disclose any information you provided on either of those forms that you believe is relevant to this group.

We especially are interested in grants, research or consulting related to the work today as well as being involved with any of the measures specifically for this cycle that we're evaluating.

So just a few reminders. You sit on this group as an individual. You do not represent the interests of your employer or anyone who may have nominated you for this committee.

We are interested in your disclosures, both paid and unpaid activities, that are relevant to the work in front of you.

Finally, just because you disclose does not mean that you have a conflict of interest. We do verbal disclosures in the spirit of openness and transparency.

Now we'll go around this virtual table. So you can see the list of names on the screen here.

I'll start with our committee co-chairs. So I'll call your name. So when I do so, please state your name, what organization you are with and if you have anything to disclose.

If you do not have any disclosures, please just state that "I have nothing to disclose," to keep us moving along.

If you experience trouble unmuting yourself, please raise your hand so that the staff can assist you with that.

Okay. So I'll start at the top and go from the left column and then the right. So starting with David Nerenz.

Co-Chair Nerenz: Dave Nerenz, Henry Ford Health System, Detroit, nothing to disclose for this meeting.

Dr. Pickering: Great. Thank you, Dave.

Co-Chair Teigland: Hi. Christie Teigland. I am with Inovalon, and I have nothing to disclose.

Dr. Pickering: Thank you so much, Christie.

And I believe Matt Austin is not going to attend this meeting, but I'll just check in just in case. Matt Austin? Matt Austin.

Okay. John Bott?

Member Bott: Yeah, John Bott. I'm an independent contractor. I currently provide contracted services for the alliance in Wisconsin, the Leapfrog Group.

I did some -- a little consulting a few years back for Yale-CORE related to interpretation of draft federal regulations. Thanks.

Dr. Pickering: Okay. Thanks, John. And nothing to disclose for the measures under review today.

Okay. And then Daniel Deutscher.

Member Deutscher: Hello. This is Daniel. I am with the Net Health systems in the U.S. and Maccabi Healthcare System in Israel, and nothing to disclose today.

Dr. Pickering: Great. Thank you so much.

Marybeth Farquhar? Marybeth Farquhar?

Okay. Jeffrey Geppert?

Member Geppert: Jeffrey Geppert from Battelle and nothing to disclose with respect to the measures under discussion today.

Dr. Pickering: Thank you so much.

Larry Glance?

Member Glance: Good morning. I'm from the University of Rochester and I also have nothing to disclose relevant to the measures under discussion for today.

Dr. Pickering: Thank you.

Joseph Hyder? And he may not be here as well today. Joe Hyder?

Okay. Sherrie Kaplan? Sherrie Kaplan?

Okay. Joseph Kunisch?

Member Kunisch: Good morning. Joe Kunisch with Harris Health System and I have nothing to disclose.

Dr. Pickering: So Joe, I think we had you having some recusals for two of the measures -- or three of the measures this cycle; is that correct? Cesarean birth and --

Member Kunisch: Yes. I think it was the perinatal ECQMs. I had been part of the advisory committee on those and I think we did some testing on one of those measures also.

Dr. Pickering: Okay. Great. Thanks, Joe. So for those measures which were the perinatal, as you mentioned, that's 0471e, 0716e and -- if I could find the other one -- 3687e, we would ask that you would be recused from discussions and voting on those measures because of the involvement with the testing.

Member Kunisch: Okay.

Dr. Pickering: So just confirming that.

Member Kunisch: Yes.

Dr. Pickering: Great. Thanks, Joe.

Okay. Paul Kurlansky?

Member Kurlansky: Kurlansky, yeah. Columbia University. I sit on the Quality Measurement Task Force of the STS, but I don't think any of those proposals are before the committee today. So I have no disclosures.

Dr. Pickering: Great. Thanks, Paul. And we also confirmed you didn't mention any disclosures or conflicts in your form. So thank you.

Okay. Zhenqiu Lin?

Member Lin: Yeah. Hi. This is Zhenqiu Lin from Yale-CORE, and Yale-CORE collaborate with The Joint Commission on the development of measure 3687e. So I will be recusing myself from that measure discussion.

Dr. Pickering: Great. And Zhenqiu, I think we also have you on for 2377, being recused for that. That's the overall defect-free care for AMI.

Member Lin: 2377, oh, okay. Yeah. I saw that this one was not -- will not be discussed in the -- in today and tomorrow's meeting, right?

Dr. Pickering: We didn't have -- so looks like we didn't have you recused. So what was the measure number that you said you were recused for? We have you listed as 3613e.

Member Lin: 3687e.

Dr. Pickering: 3687e. No, we didn't have you listed for that. We had you listed for 2377.

So just having the team confirm that real quick, Zhenqiu, I'll circle back with you. I want to keep going. We'll just circle back to just confirm this really quick.

Member Lin: Okay.

Dr. Pickering: Okay. So Jack Needleman?

Member Needleman: Good morning. Jack Needleman, UCLA, nothing to disclose.

Dr. Pickering: Great. Thank, Jack.

Okay. Eugene Nuccio?

Member Nuccio: Good morning. Gene Nuccio, University of Colorado, Anschutz Medical Campus, nothing to disclose. Thank you.

Dr. Pickering: Thank you.

And Sean O'Brien?

Member O'Brien: Good morning. Sean O'Brien from Duke University. Nothing to disclose.

Dr. Pickering: Thank you.

Jennifer Perloff? Need to stay on mute for the next 25 minutes. Sorry, Jen. Sorry, was that you?

Member Perloff: Yes. I actually am good now. I'm here and still no conflicts.

Dr. Pickering: Great. Thanks. Okay.

Patrick Romano?

Member Romano: Yes. Hello. This is Patrick Romano from UC-Davis Health in Sacramento, California.

For this meeting, I am recused from measure 2820 where I've worked as a consultant to the measure developers at UCSF.

I also worked extensively with AHRQ and CMS on risk-adjusted outcome measures, but none of those measures are under discussion today.

Dr. Pickering: Okay. Thanks, Patrick. And just confirming that, yes, the 2820, the pediatric computed tomography radiation dose measure, being associated with that. So recusing you from those discussions today.

Member Romano: Right.

Dr. Pickering: Or tomorrow. Excuse me, tomorrow. Great. Thanks, Patrick.

And Sam Simon?

Member Simon: Good morning, everyone. Sam Simon. I'm with Mathematica, and no disclosures for today's measures.

Dr. Pickering: Great. Thank you.

Alex Sox-Harris? Alex Sox-Harris?

Okay. Ronald Walters?

Member Walters: Ron Walters, University of Texas MD Anderson Cancer Center. I have nothing to disclose and no conflicts.

Dr. Pickering: Great. Thanks, Ron.

Terri Warholak?

Member Warholak: Good morning. Terri Warholak, University of Arizona. I have nothing to disclose and no conflicts.

Dr. Pickering: Thank you, Terri.

Eric Weinhandl?

Member Weinhandl: Hi. Eric Weinhandl, Satellite Healthcare, outpatient dialysis provider based in California.

I am recused from measures 3679 and 3697, the two that were submitted by the Kidney Care Quality Alliance, due to my participation in a workgroup that developed those measures.

Dr. Pickering: Great. And we also have you for 3696 as well, Eric. That's the standardized modality switch ratio for incident dialysis patients.

Member Weinhandl: Yeah, I was on the CMS team where that measure was presented and discussed. I don't know if that constitutes necessitating my recusal.

Dr. Pickering: So it was listed that -- I'm sorry, you had said that -- what was the nature of the involvement, Eric, for that measure?

Member Weinhandl: I was on a CMS task force where that measure was presented to us by the measure developer.

Dr. Pickering: So we had you listed as being recused on there. It was determined that you would be recused from that measure, so both for --

Member Weinhandl: Okay.

Dr. Pickering: So just circling back, so 3679, 3696 and 3697 being recused.

Member Weinhandl: Okay.

Dr. Pickering: Yeah, and again, the 3696 not being discussed.

Okay. And lastly, Susan White.

Member White: Susan White here from Ohio State University Wexner Medical Center. I have nothing to disclose at this time. Thank you.

Dr. Pickering: Thank you so much, Susan, and nothing to disclose.

And Zhenqiu, just coming back on you, yeah, so the

3613e was a related measure to the 2377.

So this is why there was some recusals there for 2377. So that was what was the recusal piece there, just following back on that, okay?

Member Lin: Okay.

Dr. Pickering: Anyone else join late that I didn't recognize on today's call?

(Pause.)

Dr. Pickering: Okay. All right. So well thank you very much. So I'd like to let you know that if you believe you have or might have a conflict at any time during the meeting as topics are discussed, please feel free to speak up.

You may do so in realtime during the web meeting, or you can send a message via chat to your chairs or to anyone from the NQF staff.

If you believe that a fellow committee member may have a conflict of interest or is behaving in a biased manner, you may point this out during the meeting, send a message to the chairs or to NQF staff.

Does anyone have any questions or anything they'd like to discuss based on the disclosures presented today?

You can use the raised hand feature or the chat as well. You can take yourself off mute.

(Pause.)

Dr. Pickering: Okay. No questions. So we do have 19 of our members present today. So it looks like we have quorum.

So we will maintain that, but please if you have to step away at any point in time, please let us know.

You can message us through the team's chat directly on -- we prefer that, or you can potentially send an email, we'll be monitoring that as well, but just let us know if you're stepping away and when you might be back.

But if there's no other questions, I will continue to move through our agenda today, and I'm going to turn it over to my colleague, Hannah Ingber, to walk us through the meeting overview.

Meeting Overview - Hannah Ingber

Ms. Ingber: Thanks, Matt, and good morning, everyone. Yeah, I'll go over the meeting overview, which shouldn't differ too much from our past few meetings.

Next slide, please. So we've already completed the welcome, introductions and disclosures of interest. Thank you, everyone.

Next, Gabby will give some evaluation updates from the last cycle and this one. We'll then go over our usual process overview and evaluation reminders, which are consistent throughout the -- from the past cycle.

We'll then go into the Spring 2022 measure evaluations. We'll have a break between 12:50 and 1:20 today and then we'll continue with five more measure evaluations after that.

Then at the end of the day, as always, we'll have an opportunity for NQF members and public comment; and we'll go over some next steps before Day 2 tomorrow and adjourn until the next day.

Next slide, please. So we just want to remind you that -- of a couple more meeting ground rules.

There's no rank in the room. This is a shared space. Every voice is important, and we want to emphasize that each member holds equal value on this call and in the broader scope of the work.

As NQF staff, we do our due diligence to encourage panel members to adequately review the measure information prior to the evaluation meeting.

And today, we invite you to remain actively engaged and cognizant of the varying experiences of those on the call.

Please remember to allow others the space to contribute and keep your comments concise and focused on the criterion at hand.

As always, we look forward to learning from everyone on the call today, as usual.

Next slide, please. So there's a couple meeting materials that we always use. The main one is the discussion guide, which is a synopsis of the scientific acceptability content, and it also contains Appendix B, which is additional information provided by measure developers.

That's posted online on the website and you should all have that supplied to you last week.

We also rely a little bit on the information in these background materials that are other NQF reports, including the Testing Task Force, the Measure Evaluation Criteria and Guidance from 2021, and the SMP Measure Evaluation Guidance, which is a sort of guide to scientific acceptability in particular.

Next slide, please. All right. I'll now pass it to Gabby for the Fall 2021 evaluation updates. Thanks, everyone.

Evaluation Updates - Gabrielle Kyle-Lion

Ms. Kyle-Lion: Good morning, everyone, and thank you, Hannah. Like she said, I will be going over the Fall 2021 evaluation updates and giving an overview of our current Spring 2022 cycle.

As a reminder, in Fall 2021 there were 12 measures that were evaluated by the SMP. Of those 12 measures, 7 were discussed at the meeting that happened in October.

Eight of the 12 total measures evaluated by the SMPs passed and then were further evaluated by their respective standing committees.

Of those eight measures, two were consensus not reached by the SMP and had to be re-voted on by the standing committees, and one did not pass standing committee evaluation. And that information is further talked about on this next slide here.

The two measures that were consensus not reached by the Standing Committee -- or sorry, by the Scientific Methods Panel was 3667, Days at Home for Patients With Complex, Chronic Conditions -- the SMP passes on reliability and it was CNR in validity.

The Standing Committee did not recommend this measure for endorsement due to concerns with the measure's validity, specifically their risk-adjustment model.

Measure 0689 was also a pass on reliability and a CNR on validity. the Standing Committee did recommend this measure for endorsement.

They were ultimately satisfied with the developer's response to the risk-adjustment issues and gave recommendations on how to test that model further.

This next slide shows the performance metrics table over the past few cycles that the SMP has existed.

We did want to note that these slides have been updated a little bit further, so the numbers in the "Percent agreement" column will probably look a little bit different than the PDF copy that you all have received.

But as a preference to reviewing this table in October when we showed you this, the feedback was to more accurately reflect percent agreement calculations.

And so for this meeting, we revisited each of these cycles and removed the consensus not reached measures, measures that were withdrawn, and measures that failed at the Standing Committee for criterion other than -- must pass criterion other than reliability and validity.

So you'll see that in Fall 2017 of the four measures that ultimately were reviewed by the Standing Committee, all four agreed with the SMP's recommendations.

In Spring 2018, there were six measures that were reviewed by the Standing Committee; however, that -- there were two that did not pass because of the Importance to Measure and Report Criteria.

And because of this, those two are removed from the denominators. So the percent agreement for this cycle is three out of four, or 75 percent.

In Fall 2018 of the 25 measures reviewed by the Standing Committee, 24 passed or were in agreement with the SMP's decisions.

In Spring 2019, we were able to find out more information on the one measure that did not pass and it didn't pass the Importance to Measure and Report Criteria.

Therefore, we removed it from the denominator, so the percent agreement is 39 out of 39 or 100 percent. Previously, on your slide deck, it would say 39 out of 40.

In Fall 2019 of the 15 measures reviewed by the Standing Committee, 14 were in agreement with the SMP's decisions.

In Spring 2020 of the 15 measures reviewed by the Standing Committee, 12 were in agreement with the SMP decisions.

In Fall 2020, we had originally had that of the 19 measures that were reviewed by the Standing Committee, 18 were in agreement; however, we found that there were actually two measures that did not pass of the 19 and they failed on the evidence criteria.

So because of this, the actual percent agreement here is 17 out of 17 or 100 percent.

In Spring 2021, your slide deck will say that of the 20 measures reviewed by the Standing Committee, 19 were in agreement with the SMP; however, after reviewing further, we found that two measures did not pass on evidence. And the third passed the Standing Committee and CSAC review, but is currently undergoing appeal and should not be part of the decision -- the calculation. Therefore, the true percent agreement for Spring 2021 is 18 out of 18 or 100 percent.

And then in Fall of 2021 of the six measures that were reviewed by the Standing Committee, all six were in agreement with the SMP decision. So six out of six or 100 percent.

And so far for Spring 2022, based on the preliminary analyses done by the SMPs, there were five measures that passed, one that did not pass, and seven where consensus was not reached and the percent agreement is obviously yet to come.

I'll now move into the Spring 2022 cycle overview. So there were 13 complex measures assigned to SMP. 10 of those 13 measures were new.

There were two subgroups with 12 SMP members each that were -- each subgroup assessed six or seven measures.

Of the 13 measures, five passed reliability and validity. Seven measures were consensus not reached on reliability or validity. Of those three -- of those seven CNR measures, three did not pass on reliability. One measure overall did not pass on validity and reliability.

There are four measures slated for a re-vote. Though if the SMP chooses, they can vote -- re-vote on the other six measures up for discussion, and like I stated, there are ten measures slated for discussion.

And again, of the 13 measures, eight were outcome, one was a composite, and four were intermediate clinical outcomes.

And Matt, I believe you wanted to add a point of clarification regarding measure categorization. So I will let you jump in.

## Process Overview and Evaluation Reminders - Matthew Pickering and Hannah Ingber

Dr. Pickering: Yeah. I'll just quickly mention -- I know that there's been a lot of SMP comments related to measure categorization as far as outcome versus process.

Just a quick note here is that, you know, these categorizations of measure type are largely coming from the developer. And so this doesn't always necessarily reflect what the SMP should think.

So keeping that in mind, to evaluate the measures as they have been presented by the developer.

Some of the discussions we'll have today may circle back on this topic, but the issues around outcome versus process can also be considered by our standing committees, especially if there's stronger clinical rationale involved in some of those categorizations.

All this to say is that these are largely from where the developer has categorized their measures, and it doesn't necessarily mean this is what the SMP should think.

So out of -- what comes out of the SMP, please note that we're not saying the SMP thinks this is an outcome or a process measure, but it's being evaluated as such because it's being presented to the

SMP as such.

Some of those other decisions, especially with the clinical rationales to support those categorizations, would also reside with our standing committees as well.

So just wanted to make a note of that because it was an area of discussion within this SMP for some of the measures as well.

Back to you, Gabby.

Ms. Kyle-Lion: Thanks for that clarification, Matt. And then just one more slide before I pass it back to you, Matt.

These are the 10 measures that are slated for discussion. You'll see in Subgroup 1 there are four measures, and in Subgroup 5 -- or sorry, in Subgroup 1 there are five measures, and in Subgroup 2 there are also five measures.

The stars next to them signify if they will be re-voted on because consensus was not reached, if they will be discussed and potentially re-voted on because the developer submitted additional information, or if a measure was pulled for discussion by the SMP.

All right. If there are no questions, I will pass it back to you, Matt.

Dr. Pickering: Thank you, Gabby. Let's see if there's any questions.

Member Needleman: Just a quick question, Gabby. In the discussion guide, we've got the additional comments from the developers.

Have any come in since on any of the measures that were pulled, or is the discussion guide what we have from the developers?

Ms. Kyle-Lion: Hannah or Matt, you can correct me if I'm wrong, but I believe all information is included in the discussion guide.

Dr. Pickering: That's correct.

Member Needleman: Okay. Thank you.

Dr. Pickering: Thanks, Jack.

Okay. Okay. So we'll continue on to the next session here. So now we want to talk a little bit about process overview and evaluation reminders. So just sort of reminders as we're getting into these evaluations.

Sorry, next slide. Thanks, Gabby. We have four overall ratings. We have a High rating. So this is assigned to those measures that have accountable entity level testing. So when that's required. Some of this is required for measures.

So we -- the highest rating you can get for accountable entity is a High. So there are additional considerations that could drop that High down to a Moderate.

So just keeping that in mind, may be eligible for a High, but the sampling method or results may warrant a Moderate rating.

All right. So then for the Moderate rating, the highest eligible overall rating for this one is for patient and encounter level testing only. And also for face validity, if that is conducted for a new measure specifically, right?

So this is a Moderate rating. So they've only done patient-level or encounter-level testing.

The highest rating you can get is a Moderate on this. Or if it's face validity, that's also a Moderate rating.

But again, your decisions may change, you may drop down to a Low if there's issues with -- that you find with the measure or you have concerns with the testing.

So another example there is that the sampling method and results may warrant a Low rating, for example.

And moving to Low, this is used primarily if testing results are not satisfactory or there is inappropriate methodology applied.

Inappropriate methods or the results are not satisfactory, you get a Low rating for those.

And an Insufficient is reserved if there's not enough information -- or not sufficient enough information to assign one of the other three categories.

So there's unclear specifications, unclear testing, not conducting criteria as required.

Going to the next slide, I'm just resurfacing the quorum component here that a meeting quorum is met with 66 percent of active SMP members in attendance.

And so achieving consensus is calculated from the percent of that quorum number in attendance during voting.

So those eligible participants in eligible meetings, those who are not recused from those measures.

So with the SMP scientific acceptability, the evaluation results here. So a Pass or a Recommend, this is when greater than 60 percent of Yes votes come in for those who are eligible to vote and are voting.

All right. So those being present on the call who are eligible, those being not being recused from the measure, if greater than 60 percent are Yes votes, that is a passing vote.

Consensus not reached, that's the threshold of 40 to 60 -- inclusive of 40 to 60. So if 40 to 60 percent of the votes come in -- are Yes, it is a consensus not reached.

And then for not passing or not recommending, it's less than 40 percent of the Yes votes.

So less than 40 percent of those Yes votes -- or less than 40 percent of the overall votes being Yes is a Not Pass or Not Recommended vote.

Next slide. So just some differences in the testing requirements by measure type. So for health outcome measures, intermediate, clinical outcomes, cost/resource use, structure, process, both reliability and validity, NQF requires either the patient- and encounter-level testing or the accountable entity level testing. So more of that measure score level testing.

Both types are preferred, yet currently -- yet not currently required. So not having both required, either one is sufficient.

So this can impact the rating as previously described. So depending on any concerns being raised or any issues you find with some of the testing results.

You do have an exception with face validity as testing for that measure score level or that accountable entity level for new measures is accepted as a form of accountable entity level testing.

If a patient- and encounter-level validity testing is provided, we do not require additional reliability testing.

So if they've done it for validity -- in this case, we have some measures that have performed this.

So if validity testing is conducted at the patient- and encounter-level, it can serve as the testing for reliability.

However, when we are voting on these measures for reliability, for example, you will use the validity testing to make your assessment on your vote for reliability.

We will vote separately on validity because there are other threats to validity that need to be taken into consideration.

So we will vote on reliability using data element validity testing, if that is the case, and then voting on validity separately to consider other issues or threats to the measure.

Next slide. So just some differences in the testing requirements for instrument-based measures or those measures that are also inclusive of patient-reported outcome performance measures.

So for reliability and validity, testing is required at both the patient/encounter and the accountable entity level for initial endorsement evaluation.

For the patient/encounter-level testing, it must be conducted for reliability and validity for the multi-item scales at the patient level.

For accountable entity level testing, that must be conducted for reliability and validity testing of the actual performance score -- or performance measure at the level of analysis defined in the measure specifications.

Again, face validity testing for the computed measure score is accepted as the initial form of testing here for accountable entity level validity testing.

Okay. Next slide. And then for composite, so we provide -- we also provide guidance for composite measures.

So components of composite measures should have their own properties of reliability and validity, and NQF does not consider multi-item scales in surveys/questionnaires as composites.

And NQF does not consider multiple component measures without single performance rate and multiple component performance rates as composites.

Accountable entity level testing or reliability testing of the composite is required and demonstrating reliability of individual components alone is not sufficient to pass the criterion.

Accountable entity level validity testing is not required until maintenance, and additional scientific acceptability subcriterion is required for composite measures.

So empirical analyses supporting the composite construction including the value of the components to the composite and the component aggregation and weighting consistency to the composite quality construct. So evaluating the overall construct of the composite measure.

Next slide. So just a few other reminders. All testing must align with the specifications.

This is not a new requirement. NQF is more rigorous in upholding this requirement particularly for the level of analysis testing and minimum sample sizes.

So if multiple levels of analysis are specified, each must have a test -- each must be tested separately. So if you have clinician level versus facility, each must be tested separately.

NQF's requirement permits passing some levels and not all. So if one level looks sufficient, but another doesn't, we can pass the measure or you may deem the measure to be suitable for reliability and validity at one level, but not another.

Occasionally there are several performance measures included under one NQF number as well. Each measure must be evaluated separately.

So sometimes you have different levels under the same number, and each must be evaluated separately. So some measures may pass, again, and others may not.

Okay. Next slide. In the consideration of risk adjustment -- so you've seen some of this come through email communications in advance of this meeting. This is not a new slide.

We've definitely talked about this previously as well, but for risk adjustment, this is -- this assessment of risk adjustment is required for all outcome, resource use and some process measures where there's justification.

So inclusion or exclusion of certain risk factors in the risk adjustment approach should not be a reason for not passing a measure.

Rather the concern should be focused on some of the discrimination or calibration statistics, or the overall approach and method of adjustment.

Those are grounds for not passing a measure whereas exclusion of those of certain factors based on the clinical aspects or even what may be in the provider's realm of influence, locus of control, if you will, to impact a certain factor, that is more so a conversation for our standing committees. We really are looking for the SMP to evaluate just the overall approach from the calibration statistics and discrimination.

And I know that sometimes there's a gray line there, but if we could try to fall back on those two bullets right there the best we can during those discussions, that would be very helpful.

And any of those recommendations or concerns, keep in mind, will be communicated to the standing committees in their evaluation to the assessment of these measures.

So in the absence of risk adjustment for an outcome, resource use or even some process measures, a strong rationale should be provided by the developer for consideration.

For all measures, incomplete or ambiguous specifications are also grounds for not passing a measure.

The empirical validity testing is required at the time of maintenance evaluation. If that is not possible, strong justification is required and must be accepted by the Standing Committee.

Next slide. So the SMP also has articulated other guidance for submissions in the past. We're just going to touch on those.

So that being provide greater detail when describing the testing methods and results -- sorry, I think we might have someone speaking off mute.

Okay. Thank you. We also -- the SMP has also asked that developers also provide an overall statistic when conducting signal-to-noise reliability testing to provide a variation around that overall central tendency.

And then also provide greater detail in describing the construction of validity as testing as well.

So what are the hypothesized relationships, why are these relationships an indicator of validity, as well as the expected direction and strength of those associations with that testing.

And then in the results is to specify the results and the interpretation of how that relates back to the hypothesis and what was expected for the validity testing, but lack of No. 2 and 3 here should not be grounds for not passing a measure.

I mean, these are very nice to have, but currently not providing a variation around some sort of central tendency for your overall statistic, as well as not providing some of this detail, is not grounds for not passing, but it is something we can definitely take as concerns or recommendations for the Standing Committee's consideration.

Okay. Next slide. All measures reviewed by the SMP can be discussed by the Standing Committee. So this is now after our meeting today.

So after the votes have been done, the tallies have been made, all these measures can be discussed by the Standing Committee.

Standing committees will evaluate and make the recommendations for endorsement for those measures that passed from the SMP review on reliability and validity, and those measures where consensus was not reached.

So the Standing Committee will have to re-vote on those criteria that there's still a consensus not reached decision.

So measures that do not pass SMP, either on reliability or validity, may be pulled by a Standing Committee member for further discussion and re-vote if it's an eligible measure.

And so what does "eligible" mean? We'll go to the next slide for the criteria of that as a reminder.

Measures that did not pass due to the following reasons are not eligible for a re-vote. They can be pulled for discussion, they just would not be eligible for re-vote by the Standing Committee: Those with inappropriate methodology or testing approaches, incorrect calculations or formulas that have been used, description of the testing approach or results or even data are not sufficient for SMP to apply the criteria, and the appropriate levels of testing are not provided or otherwise did not meet the NQF's minimum evaluation requirements.

Okay. Any questions there before we continue?

I also see that Alex Sox-Harris has joined -- thank you, Alex -- from the VA and Stanford. He has nothing to disclose for the measures today, and he'll be listening and beginning, you know, unmute yourself as we proceed. So thanks, Alex, for joining.

Okay. With that, I'll turn it back to Hannah to go through our voting process. Hannah?

Ms. Ingber: Thanks, Matt. Yes, so nothing has changed here from prior cycles either, but just as a reminder -- oh, we have a question?

Member Romano: Can you hear me?

Ms. Ingber: You're coming in a little bit fuzzy. Was that Patrick?

Dr. Pickering: I think it was, Hannah. We may try to message him. See what's going on with his audio.

Ms. Ingber: Okay. While we troubleshoot that --

Member Romano: I --

Ms. Ingber: Oh, go ahead.

Dr. Pickering: Sorry, Patrick. If you can hear us, we're hearing bits and pieces of your audio, but I'm not sure if you're trying to ask a question or --

Member Romano: Can you hear me?

Dr. Pickering: Not very well. Not very well. Patrick, we'll circle back with you and see if we can get your audio squared away.

Member Romano: Okay. Is this better?

Dr. Pickering: Yes.

Member Romano: Okay. This is better now?

Dr. Pickering: Yes, it is.

Member Romano: Okay. Okay. Thank you.

I just had a question on the last slide. I assume that -- this one here, yes. So in terms of the committee consideration of measures that do not pass the SMP, I assume you're referring specifically to the results of this meeting, not -- right, not the subvotes that occurred prior to this meeting, correct?

Dr. Pickering: That's correct. It can also apply to the subvotes prior to this meeting, especially if there's -- the developer did not provide a response to those.

We wouldn't be discussing them and re-voting on those because those subvotes sort of stand because the developer has not provided any response to those.

But for those that did provide responses, we will be considering those today. So those measures that did not pass in the subgroup preliminary votes, the developer did provide responses.

We will be discussing those measures and the outcome of those, yes, if they still do not pass, there still may be some decision-making that could occur whether or not they're eligible for re-vote by the Standing Committee.

Member Romano: Okay. Thank you.

Dr. Pickering: Great. Thanks, Patrick. We can hear you a lot better now. Thank you.

Sorry, Hannah, go ahead.

Ms. Ingber: Thanks. So the measures discussed by the SMP today were determined during those SMP measure review activities that we were just alluding to, the preliminary analyses.

So first, staff will briefly introduce the measure and the testing that was provided, and then SMP member lead discussants, which are noted on each of the slides, will summarize the key concerns from the SMP and any other details from their preliminary evaluation.

Then other subgroup members are invited to comment on the measure and then developers are given about two to three minutes for an initial response and may respond to the SMP's questions directly as well.

We'll then open the discussion one more time to the full SMP and proceed in voting by the individual criterion.

And as a reminder, those members who are recused cannot discuss measures where conflicts are identified, as noted at the beginning of this meeting.

Next slide, please. Thanks. So voting is conducted synchronously, virtually and confidentially via Poll Everywhere.

We sent a link this morning with that same link to the SMP members and voting will occur following each criterion discussion.

So only the SMP subgroup members are voting on measures that they were assigned, and again, recused SMP members can't vote for measures where conflicts are identified, but none of the subgroup members have an overlap in that way.

So then the subgroup voting results will be taken during the meeting and are the official SMP vote.

Any other measures that are not pulled for discussion will pass in a consent calendar vote from the preliminary results.

Next slide, please. I will now pass it to Gabby to do the voting test.

Voting Test

Ms. Kyle-Lion: Hello, everyone. Just give me one moment to share my screen.

Okay. So like Hannah mentioned earlier, you should have gotten a link to the voting poll via email, I think, yesterday and this morning.

If you do not have access to that link, please let Hannah or I know and we can send that to you.

The test question today is: Do you like Brussels sprouts? Your answers are A for yes, or B for no. And I believe we have 20 members on the call, so we're looking for 20 votes here.

(Pause.)

Ms. Kyle-Lion: Sorry, I just want to clarify that this is just for the Scientific Method Panel members and only them.

Nobody else should be participating in this vote, so just the 20 Scientific Method Panel members.

(Pause.)

Dr. Pickering: If you really don't like them, can you vote two or three times?

(Laughter.)

Ms. Kyle-Lion: I wish.

Dr. Pickering: They're so good, though. Crispy sometimes with a little bit of bacon. It's very good. Good stuff. Get a well-ventilated kitchen though, if you're cooking them.

Ms. Kyle-Lion: And it looks like we're at 17 votes and we need 20.

Dr. Pickering: Is anyone having difficulties with the Poll Everywhere link?

Ms. Kyle-Lion: Hannah, I think Paul said that he's having issues. Could you possibly resend him the link?

Ms. Ingber: Yes. Happy to do so.

Member Kunisch: Could you send it to me too, Joe Kunisch, also? I'm having trouble finding that link.

Dr. Pickering: Sure. Thanks, Joe. We'll get that to you. So, that's two individuals. Anyone else?

And then Alex, looks like, having some unrelated computer issues may be preventing his vote. So, that would be a total of 20 with those three individuals.

So, I think we're just confirming -- anyone else having any issues? Okay. I think with Alex, Joe and -- I can't remember the third. My apologies. I think that makes 20, Gabby.

Ms. Kyle-Lion: Yes, I believe so. If we're comfortable, we can move forward if Joe and Paul and Alex are confident that later they'll be able to vote or they can also message me their vote directly, if that's needed.

Dr. Pickering: There's 18, 19 --

Ms. Kyle-Lion: Okay. I'm assuming, then, that that last vote would be -- would probably be one of those three. So, at the time when we have to vote again, you can send me your vote directly.

So, I'll go ahead and lock this poll. We have 68 percent of people saying, yes, they like Brussels sprouts; and 32 percent say no.

So, thank you so much for participating and I will pass it back to you, Matt.

Dr. Pickering: Thanks. So, it looks like we got consensus there that the measure of Brussels sprouts moves forward. Okay. Bad joke.

Alright. Well, thank you all very much. So, we'll go to the next slide there, Gabby.

So, before we go into the Spring 2022 measure evaluation, just want to see if there's any other questions or comments from the group today.

And just to confirm, Sherrie Kaplan and Marybeth Farquhar, have you joined?

(Pause.)

Dr. Pickering: Okay. Not seeing any questions in the chat or hands raised, I think we can start going through our first measure.

We'll go to the next slide. Okay. So, just a reminder for folks, NQF staff will be presenting the measures.

We'll just present any of the relevant testing information, noting the results as well. Any of the key concerns will be summarized by our lead discussants. You can see those individuals listed on the slides for convenience.

I also want to make mention that the first series of measures here are actually Subgroup 2, so not Subgroup 1. So, my apologies for that typo, but it's Subgroup 2, not Subgroup 1. So, we're starting on Subgroup 2 measures.

Dave Nerenz will be our facilitator here for our co-chair for these measures. And before we get started, I just wanted to check in to see if the developer, UM-KECC, are you on the call?

University of Michigan, anyone from University of Michigan on the call?

Dr. Shahinian: Yes, we are on the call. This is Vahakn Shahinian from the transplant urologist unit of UM-KECC.

Dr. Pickering: Great. Thank you so much. So, we will provide an opportunity for the developer to provide any remarks based on the SMP concerns as well as they are here for any questions as needed.

We kindly ask that developers remain -- do not chime in until called or recognized by our co-chairs. So, just want to keep that in mind.

## 3689 - First Year Standardized Waitlist Ratio

We'll go ahead and get started. So, again, Subgroup

No. 2. So, the first measure up for discussion is 3689. This is the First Year Standardized Waitlist Ratio or FYSWR.

This is a new measure and you can see the measure developer being University of Michigan Kidney Epidemiology and Cost Center, what we've called as UM-KECC.

The measure did pass reliability, but it was consensus not reached on validity. So, the discussions in re-voting today will be on validity, but I'll just note that this measure tracks the number of incident patients in a practitioner group who are under the age of 75 and were listed on the kidney or kidney-pancreas transplant waitlist or received a living donor transplant within the first year of initiating dialysis.

For each practitioner group, the First Year Standardized Waitlist Ratio is calculated to compare the observed number of waitlist events in a practitioner group to its expected number of waitlist events.

The First Year Standardized Waitlist Ratio uses the expected waitlist events calculated from the Cox model, adjusted for age and patient comorbidities at incidence of dialysis.

For this measure, patients are assigned to the practitioner group based on the National Provider Identifier, NPI, Unique Physician Identifier Number, UPIN, information entered on the CMS Medical Evidence 2728 form.

This is an outcome measure, or classified as an outcome measure. The data source is using claims and registry data. And the level of analysis here is the clinician group or practice.

It is risk-adjusted and, for the reliability testing, the SMP preliminary analysis deemed this as a Moderate rating.

The developer conducted accountable entity level testing and calculated an inter-unit reliability value of 0.64 for the measure, which indicates about 64 percent of variation, and the measure can be attributed in between-facility differences and about 36 is within-facility differences, but, again, won't be focusing on any discussions on reliability as it did pass with a Moderate rating.

For validity testing, this came in as consensus not reached from the preliminary assessments from our subgroup -- Subgroup 2.

The developer tested the measure by evaluating the association of this measure -- this dialysis practitioner group level measure performance and subsequently mortality and overall transplant rates among all patients attributed to those practitioner groups.

They examined the Spearman correlation between the practitioner group measure value and each of the outcomes, respectively.

And the practitioner group level second-year average mortality rates are 15.3, 15.7 and 15.9 deaths for 100,000 patient years for T1, T2 and T3, those tertile rankings, respectively. And then the Spearman correlation coefficient is negative 0.02.

The dialysis practitioner group level second-year average transplant rates are 4.7, 3.2 and 1.8 transplants per 100,000 patient-years associated with that T1, T2 and T3, respectively, those tertiles, and the Spearman correlation coefficient for this was 0.32.

And so, the developer noted that higher FYSWR performance correlated with higher second-year transplant rates with clear separation of transplant rates across practitioner group tertiles of performance.

And the direction of the relationship with mortality was as expected with numerically lower mortality with higher performance of the FYSWR measure, though it did not have statistical significance.

You can see some questions here about additional clarifying information from the developers.

The developer did provide some responses to some of the SMP concerns, and are there any concerns about the reliability/validity testing methodology, specifically the validity testing in this case as we are re-voting on that.

Our lead discussant here is Ron Walters, but I'll turn it back to Dave to facilitate the discussion.

Co-Chair Nerenz: Thanks. That's a good summary and I'll turn it around in just a second.

I just would suggest, as we entered this whole set of measures for Subgroup 2, and remember these are quite similar measures to each other, conceptually related, same underlying dataset used in a lot of the testing. So, some of the issues are going to be the same as we move from measure to measure.

And as we do that, let's just try to keep in mind if we have an issue that gets discussant-resolved early in the discussion, let's not do it over again de novo as we move along to the extent we can.

And also, I think I just mentioned that in the set of measures, not necessarily this one specifically, there may be a couple misunderstandings in the earlier back-and-forth about our preliminary votes in the discussion and I'll highlight a couple of those just to see if we maybe get them off the table.

One is in the issue of use of patient-months. The SMP, as I follow the discussion, was not concerned about use of patient-month as the vehicle for calculating numerator/denominator. It's been done before in other measures. Nothing wrong with it.

The concern was in whether the nonindependence of patient-month observations was appropriately accounted for in reliability and validity calculations. So, we got to make sure we stay focused on that.

And then similarly there was a discussion, for example, of use of beta-binomial methods as a way of establishing reliability.

We had no concern about that as a method in general. Again, it's fine. The question is, are the underlying assumptions met given the particular set of observations.

So, I'll turn it over to Ron at this point; but since we have a lot of ground to cover, just want to make sure that we don't get off into areas that actually there's no real dispute about.

Okay. Thanks. Ron, all yours.

Member Walters: I was going to start out exactly the same way that you're going to see duplication across these -- and learn to love the term "waitlist" very dearly.

So, I'm trying not to be terribly duplicative. Again, we won't talk about reliability. Validity is the focus of the day.

And I'm going to hit the nail on the head at the beginning from the discussions that went on. Yes, I think, the intro by Matt, we should consider this as an outcome and that's what the -- that's what the measure developer submitted it as, and there's certainly a lot of people who felt strongly as a process measure.

I have contemplated that ever since the discussion erupted. I contemplated it while I was reviewing the measure and I think that the -- in the measure developer's mind, again, and if you're probably the patient, a case can be made that, yes, there's a lot of processes that go into getting on the waitlist, but being on the waitlist is a very significant intermediate clinical outcome, I would say. Because if you don't get on that waitlist, you aren't going to get a transplant. So, I think we should consider that in our discussions about all of these measures.

The risk-adjustment model was utilized and, as you heard, the probably somewhat disturbing thing is that the basis of empirical testing was correlation to mortality, and that correlation ended up to be relatively weak certainly the first year, and also to getting a transplant where the correlation was a little bit stronger positively, but perhaps not as strong.

Obviously, there's a lot of factors that go into ending up getting a transplant within the first year as well as whether or not you live long enough to get that transplant. So, those were major issues as far as validity discussion.

My last comment was that of course we'll have to talk about, in this one and subsequent ones, if we view this as a process measure, then empiric validity testing kind of goes out the window and they can be for the first -- for the initial -- for a new measure, and we can make those suggestions that these are things that the measure developer consider for measure maintenance and may be important considerations for the future.

So, I think these issues are all going to be common to all the measures that have to do with waitlisting and they were brought up in all the comments that are made.

So, I won't belabor this anymore other than this has been the subject of a lot of discussion. Thank you.

Co-Chair Nerenz: Okay. I guess the next step is, are there any other comments from members of Subgroup 2 following Ron's overview?

Member Romano: I had one comment that I'd just like to perhaps engage the developers on.

The measure -- the first measure that we're discussing, 3689, which is the measure of the First Year Standardized Waitlist Ratio, has a prior missing data rate of 6.2 percent compared with about 1 percent by the other two measures.

And this missingness is apparently attributable to some difficulty matching the practitioner and I was, frankly, confused about why this would be different for the three measures. The developer responded, but I didn't understand the developer's response.

Basically, the question is: Did the IDR -- that is, the Integrated Data Repository maintained by CMS -- is used in all three measures to link a given provider to their practice group and then a 2728 is used to identify the provider.

So, this seems to be done in the same way for all three measures. I don't understand why it would be different.

The 2728, when it's missing, those records have to be excluded. That's about one percent.

So, I'll be looking for a little bit more clarification from the developers about why there's a missingness issue that's unique to 3689.

Co-Chair Nerenz: Thanks, Patrick.

Anyone else?

Member Nuccio: Dave, this is Gene Nuccio.

I'd also like the -- some more clarification regarding the decision to create tertiles as opposed to, perhaps, quintiles for the approximately 2300 or so dialysis groups that were measured.

And also, related to that is movement across the tertiles -- or quintiles, preferably, but tertiles -- from the first year to the second year.

And the third issue that bothered me was that in the first year the outcome reported was mortality rates. And in the second year, it had transplant rates.

And so, why were they changing the outcome that they were measuring across those two years?

So, the three issues are why tertiles as opposed to some more smaller grouping, still significant numbers, a quintile, perhaps? What's the movement from one year to the next for the same outcome that is for transplants or for mortality?

Co-Chair Nerenz: Thanks, Gene.

Anyone else?

Member Bott: This is John Bott. I had my hand up, but maybe you can't see it, David.

Co-Chair Nerenz: Yeah. I'm sorry. I can't see the hands up. Just jump in when you want to.

Member Bott: Yeah. There's several data elements in the risk adjustment that seems like they could have occurred before or after the onset of care that are captured on the CMS form 2728.

I'm not a Jedi on CMS form 2728, so perhaps the measure developer can address that ambiguity there. That would be appreciated. Thanks.

Co-Chair Nerenz: Thank you, John.

Anyone else?

Member Romano: I guess one final point I'll make for the record. There was some discussion, you know, in the discussion guide related to the role of social risk factors.

Recognizing what we've discussed by email and earlier this morning, it may not be the role of this committee to decide whether social risk factors belong in the model or not, but I'd like to raise that that issue should be forwarded to the Standing

Committee for further discussion.

And I personally retain my skepticism about including social risk factors in models of transplant waitlisting.

Co-Chair Nerenz: Thanks, Patrick. And I'll just say I think you've captured it just in the right way.

We can express thoughts either for or against, but ultimately it's -- we're not to be voting sort of on these substantive or content issues. Point well-taken. Thank you.

Anyone else?

I see nothing in the chat, so, let us move along now to the developer from U of M-KECC. Your turn.

Dr. Shahinian: Hi there. So, I'll start. This is Vahakn Shahinian. I'm a transplant nephrologist.

I guess I'll tackle the issue -- the first issue that was raised around the concerns about the associations with mortality and subsequent transplantation.

You know, I think as we know -- and with respect, you know, to some extent, I think this gets at our choice to call this an outcome.

I think the issue is that what is directly influenceable kind of by the dialysis practitioners is that element of getting the patient to a state of sufficient and a beneficial health status that would make them a candidate for transplantation.

But once you get there, certainly the conversion, so to speak, to a transplant is subject to many systemic and almost random factors, you know, that have to do with organ availability and that's where there's a lot of drop-off.

And I think that's why we see that the correlations, although they're in expected directions, are admittedly very modest.

But it's exactly that point for, you know, or reasoning that we focus on it as the outcome because it's what is under control of the dialysis practitioners whereas there's a lot that factors in that's kind of beyond almost anyone's control with respect to actually receiving a transplant.

And I think the associations are a reflection of that, so I think that's the main thing I'd say there.

The only other thing I can add, and we mentioned this in our response, we did not do formal face validity testing, but certainly engaging with a technical expert panel during the development process that includes a range of stakeholders that were well aware of systematic review of the literature on this topic.

They certainly demonstrated majority support for a measure that was directed at waitlisting.

With respect to clarifying why the discrepancy in the, you know, the missingness of practitioner attribution with this measure as opposed to the other measures you're going to discuss, the issue is is that the connection between a given individual physician and a practice, that's done through the IDR.

And so, that's the same with both, but the attribution of patients to that physician, that's what's different.

So, in the case of this measure, it's done through the Form 2728, whereas for the prevalent measures it is done through what's available on the Medicare claim.

And it's -- that's where there's a discrepancy between the two in terms of missingness for that attribution.

So, the way, you know, individual practitioners are attributed to practices is the same, but the way patients are attributed to those individual physicians is different and that's where that discrepancy is coming from.

With respect to the tertiles, you know, I think our choice there, you know, I guess we felt that it illustrated the relationship adequately.

I think we didn't just -- we just didn't think of cutting into finer points, I guess, is the point with that.

I am less -- I guess I'm not as clear on the second point that the reviewer was making about the movement of the tertiles.

I think, you know, our idea was to -- and we looked at this in several ways, but to look both at where practitioner group performance was with respect to the tertiles of performance and then look at the relationship with both, you know, same year, but also subsequent year, both mortality and transplant and we presented the second-year outcomes.

With respect to the question about the timing of when the data is collected on the 2728 versus when the measurement -- the performance measurement period is done, the way the 2728 is collected is it's essentially a registration form for the status of end-stage renal disease.

And so, it's effectively collected at the onset of the need for dialysis. And so, the information captured on that form essentially more or less precedes anything that comes after.

These are comorbidities, for example, that would have been present in the patient at the time they started dialysis and, therefore, by design, those factors would have been present prior to the measurement period so that -- that separation is in force that way.

You know, later on with the prevalent measures, by design, we look at comorbidities in Medicare patients in the year prior to the measurement period.

So, we do make sure that the factors that we're looking at are measured or assessed prior to the measurement period.

I will stop there. I didn't know if -- I know there was some mention early on about the issue about the nonindependence of the patient-months.

We've included a response for that, but I can also have one of my statistical colleagues comment on that as that was an issue that was raised.

Kevin, did you want to comment on the issue of the concerns around the nonindependence of the patient-months?

Dr. He: Yes.

Co-Chair Nerenz: Thank you. Just for a second --

Dr. He: Yes.

Co-Chair Nerenz: -- I'm going to turn to Ron just as essentially a little referee here.

If that issue in this measure had to do with reliability and reliability passed unanimously, we don't really need to discuss it at this point. It may come up in the next or the next measure after that.

Ron, in your mind, does this have anything to do with our validity votes or was this a reliability question?

Member Walters: It was not a reliability question, in my mind.

Co-Chair Nerenz: Okay. We need to -- this may come up in the very next measure. We're happy to talk about --

Member Walters: Yes.

Co-Chair Nerenz: -- it, but, you know, let's -- let's stay focused on what may be up for re-vote.

Ron, I don't mean to cut things off -- well, first of all, let me just ask back to our developer, Vahakn, it sounded like you were sort of winding up.

Do you have other things that we should know about or should I turn it back to Ron?

Dr. Shahinian: No, I think we can stop here. AS I think about it, the patient-month issue is really relevant for the subsequent measures. So, we can tackle that when we get there. Thank you.

Co-Chair Nerenz: That's fine. That's efficient.

Ron, anything else that you want to ask about or bring up?

Member Walters: Nothing. Nothing big. I think you heard the measure developer thoughtfully consider the kind of issues raised and this may become relevant as the morning goes on.

So, I would recommend passing this particular measure.

Co-Chair Nerenz: Alright. Let's let any other Subgroup 2 members now, after hearing from the developer, any further thoughts before we move to a re-vote?

I see Patrick's hand up, literally, in the video window.

Member Romano: Yeah. I'm wondering -- again, I'm a little confused about the missing data issue.

Could you explain why -- why is the missing data problem defined differently for this measure than for the other two measures?

I'm not understanding why you have an additional five percent of patients that have to be excluded for this measure compared with the other two.

What's the conceptual basis for that need?

Dr. Shahinian: It has to do with the -- wait. Sorry, this is Vahakn Shahinian again. It has to do with how we're attributing -- the method by which we're attributing patients to providers and, for this measure, the way we're doing that.

This is different than for the other two measures. For this measure, we're using the National Provider Identifier that's on the 2728 and there is missingness of that data element. And that's where the missingness is coming from for this measure.

Member Romano: And what's the rationale for using a different approach?

Dr. Shahinian: Oh, okay.

Member Romano: Why not use a consistent approach for all three measures?

Dr. Shahinian: The -- part of the -- the issue is that it has to do with the health insurance of the patients.

And so, the prevalent measures are measures that, by design, are limited to Medicare fee-for-service patients, and that is still a very substantial portion of the prevalent dialysis population.

And we're able to leverage the dialysis claims, the Medicare fee-for-service dialysis claims, in order to make the attribution of patients to the individual providers with the First Year Standardized Waitlist Ratio.

And, again, our rationale for this measure broadly is because we want to incentivize rapid addition to the waitlist after initiation of dialysis.

Within the first year there is a substantially lower percent of the patients that are Medicare fee-for-service patients and we wanted a measure that could capture patients of all forms of insurance within that first year.

And we, therefore, wanted to use an attribution method that was not reliant on Medicare fee-for-service claims, and so we went through the 2728 and that did, unfortunately, introduce some missingness.

Member Romano: Okay. Thank you. Got it now. Thank you.

Member Walters: Pat, yeah, I think -- I think in one of the subsequent measures, this was kind of discussed within the measure, and basically I think it's a timing issue more than anything else, who fills out the 2728.

And then as the longer time frame goes on, who submits the claims to Medicare, what's their insurance and so on and so on.

I remember reading that somewhere, but it came to a head regarding the question that you asked.

So, yeah, I think that's the issue. It's the form's different and the timing's different.

Member Nuccio: Dave?

Co-Chair Nerenz: Gene, looks like you've got a hand up.

Member Nuccio: Yeah. I'm not sure my automated hand raise works, so I used the Patrick method.

Co-Chair Nerenz: It works.

Member Nuccio: It works.

To clarify the question about, let's see, the outcome of transplant, the question -- my concern was that if a dialysis center has a high transplant rate in one -- in the first year, and then a low transplant rate in the second year, that may not be any kind of function of the dialysis center, but of the transplant opportunities that are found within geographic range of that transplant, that dialysis center.

And so, the question is, is it -- without the information about transplant in Year 1 and transplant in Year 2, I don't know whether or not that's a meaningful measure, a measure that validly addresses the quality of the dialysis center as opposed to the availability of transplant opportunities from the transplant centers geographically located.

Could the developer respond to that?

Dr. Shahinian: This is Vahakn Shahinian here, UM-KECC.

So, yeah, I mean, I think that there's a relationship or an expected to be some relationship between dialysis practitioners that do a better job at getting their patients waitlisted for them to get subsequently transplanted because of the necessity of waitlisting to proceed the, you know, even the opportunity for transplant.

We completely agree that there are other factors, like you mentioned, exactly right, that also influence transplant, which is exactly why we see, I think, perhaps a more modest correlation than you might otherwise see because there are these other -- there are these other factors.

What is under -- most under control of the dialysis practitioners is that waitlist, but, nevertheless, there is a relationship there because to the extent that more of their patients are waitlisted, there's going to be a higher probability that they will get transplanted, but there's effectively going to be this kind of drop-off or attrition because there are other factors at play, but we expected to see at least some association there because of that.

Co-Chair Nerenz: Thank you, Vahakn.

Anyone else before we move back to Hannah?

(Pause.)

Co-Chair Nerenz: Going once. Going twice. Okay. Hannah, I think we are now back in your hands.

Ms. Ingber: This cycle, I'm actually passing the mantle to Gabby.

Co-Chair Nerenz: Okay. We're in your hands.

Ms. Kyle-Lion: Alright, everyone. Okay. Alrighty. And just as a reminder, we're looking for Subgroup 2 to vote here at a minimum of eight votes, though I believe we should have ten.

So, voting is now open for Measure 3689 on validity. The options are A for high, B for moderate, C for low, or D for insufficient.

Dr. Pickering: This is Matt. As you're voting, just making note that there are no recusals for this measure. So, no recusals for this measure.

So, to the Subgroup 2 members, it's just a reminder. No recusals on this one. So, you can vote.

Ms. Kyle-Lion: I'm seeing eight votes, but I do think that we are looking for ten. So, I'll just give it one more moment.

I see nine. I'll just give it one more second to see if we get that tenth vote. And, again -- okay. We have ten. I will go ahead and lock the poll.

Alrighty. There were zero -- voting is now closed for Measure 3689 on validity. There were zero votes for high, eight votes for moderate, two votes for low, and zero votes for insufficient. Therefore, the measure passes on validity.

I will pass it back to you, Matt.

Dr. Pickering: Okay. Thank you -- make sure I'm off mute. Okay. Great. Thank you so much, Gabby.

Alright. So, that's our first measure. So, thank you, everyone. So, the measure does pass on validity.

We'll go to our next measure. And if, Gabby, if you could pull that back up on the slides?

Ms. Kyle-Lion: Yep. Just give me one second.

Dr. Pickering: Sure.

Co-Chair Nerenz: Gabby, while you're doing that -- this is Dave here. I want to thank everybody who just involved in that discussion. The panel members, the developer. That was clear, it was sharp, it was focused, it was tightly bounded, a model for how we can live today. I think that was pretty good.

Dr. Pickering: Agreed. Agreed. Thank you. Thank you, Dave, and as well as our lead discussants, SMP members and developers.

I think if we can replicate that, we will be in good shape today.

Ms. Kyle-Lion: I just want to confirm -- time out. I just want to confirm you can only see the slide, right, not the presenter view. Because I can see the presenter view, so I just want to make sure.

Dr. Pickering: I was just going to mention we see the presenter view on --

Ms. Kyle-Lion: Okay. Okay. Is that better?

Dr. Pickering: Still presenter view. That's okay. I think -- there we are. You got it.

Ms. Kyle-Lion: Okay. Perfect.

Dr. Pickering: Thanks, Gabby.

Ms. Kyle-Lion: Yep. No problem.

Dr. Pickering: Alright. Thank you. So, again, this is also a UM-KECC measure. UM-KECC is still on the call.

## 3694 Percentage of Prevalent Patients Waitlisted in Active Status

So, this is Measure No. 3694, Percentage of Prevalent Patients Waitlisted in Active Status, or aPPPW.

Again, this is a new measure. This measure tracks the percentage of patients in each dialysis practitioner group practice who were on the kidney or kidney-pancreas transplant waitlist in active status.

Results are averaged across patients prevalent on the last day of each month during the reporting year.

The proposed measure is a directly standardized percentage, which is adjusted for covariates such as age and other risk factors.

It is categorized as an outcome measure using claims and registry data. The level of analysis is the clinician group/practice. There is a risk-adjustment modeling approach that's been conducted on this measure.

And for reliability, we can see that in the preliminary subgroup assessment it did pass on reliability.

So, we won't be focusing our discussions today on reliability, but I'll quickly summarize that.

It did receive a high rating from the SMP subgroup assessments. The testing was done at the accountable entity level using inter-unit reliability and a bootstrap approach. The IUR, the inter-unit reliability, value was found to be 0.93.

I'll focus now on validity because that is where our discussions will reside for this measure in re-voting.

The developer -- you can see the consensus not reached and the developer did conduct accountable unit -- or accountable level testing at accountable entity level.

They tested validity of the measure by evaluating the association with the dialysis practitioner group level measure performance and mortality and overall transplant rates among all patients attributed to the practitioner groups.

They used Spearman correlations between the practitioner group measure value and each of the outcomes respectively.

And the dialysis practitioner group level average mortality rates are 17.8, 18.3 and 19.2 deaths per 100 patient-years for the tertiles T1, T2 and T3 respectively.

And the Spearman correlation coefficient was -0.083. Was found to be statistically significant as well.

For the dialysis practitioner group level average transplant rates, those were found to be 5.0, 4.2 and 3.1 transplants per 100 patient-years for the tertiles T1, T2 and T3 respectively. And the Spearman correlation coefficient there was 0.279, also statistically significant.

The developer did note that higher aPPPW performance correlated with higher transplant rates with clear separation of transplant rates across practitioner tertiles of performance.

And the direction of the relationship with mortality was as expected and statistically significant with numerically lower mortality and higher performance on the measure, although the magnitude of the association was smaller for transplant rate.

Okay. With that, Dave, I'll turn it back to you for our lead discussants, who is -- Zhenqiu is our primary and Eugene is our secondary, to summarize the SMP concerns related to validity.

Co-Chair Nerenz: Alright. Well, again, good summary, Matt. We'll just turn directly to Zhenqiu for a more detailed discussion focusing on validity.

Member Lin: Okay. So, for this measure, some of the same concern has been discussed in the -- prior to this. So, I will just focus on additional concern that we need to get into more detail.

I think Ron mentioned about a distinction between process measure, outcome measure.

I think the reason this is relevant is I think this will affect how people view the risk-adjustment approach for this measure, right?

If it's a process measure, you have different order in terms of what you need to adjust for.

So, the first concerns relate to the inclusion of social risk factors. I think Patrick already mention about this.

So, for this measure, the risk model include extensive list of variables. There are about 15 or 16 incident comorbidity, 64 prevalent comorbidity.

And then the variable also including ADI, Area Deprivation Index, and -- social risk factor in the model.

In addition to this, the risk model also account for transplant center characteristic.

They adjust for transplant center mortality ratio and transplant center -- transplant rate, right? And in addition to that, there's a transplant center runaway effect.

So, the question is, after all of this adjustment, what's left over? Do we think this is adequate for a measure to capture the quality of care of clinician group in taking care of this type of patient? So, this is the first concern.

The second concern is about this measure use risk-adjustment model. So, there was concern about there's no validation for risk-adjustment model.

The developer did respond and they use all the data available in the year.  I think this didn't get to the concern exactly because the model develop ideally one that could be validated.

And this model is going to be used for future reporting, right? It's not limited to the data you use for model development.

So, I ask the developer can explain a little bit more in terms of why they don't consider validation on a risk model.

The third one is about patient-month and the determination that I think David mentioned earlier.

I think multiple reviewers brought up this issue and the developer response in there refers to the material in the application form, Section B -- I think B6 or B2 or B605 -- let me see. Oh, 2B05.

I thought what's included in that section is more about collecting provider as, you know, whether or not different from the average, right?

So, because of in the study cohort about 280,000 patient, so, on average, about every patient contribute to about nine patient-month data point. So, the unit on this is patient-month.

So, it would be helpful if the developer can clarify how they address this lack of independence in their risk model.

In their risk model, four steps, you know, I for clinician group and J for transplant center, you know, K for risk factor, and then I think R is for month.

So, it would be helpful to explain how they account for lack of independence among patient-month within the same patient because it's not obvious that the patient status would change from month to month within a single year.

So, this our three main concern I have for this measure, I mean, based on the feedback from the group.

Co-Chair Nerenz: Okay. Gene, anything to add?

Member Nuccio: Along the lines of what Zhenqiu just said, first of all, I want to recognize that the question of characterizing the measure as a process measure versus an outcome measure is primarily related to the reliability.

And, in fact, the very first question we're asked to review in our evaluation is whether or not the -- let me get the right wording up here -- are the submitted specifications precise, unambiguous and complete so that they can be consistently implemented.

And that's a question that we -- that's the very first question we're asked and I would submit that the determination of whether it's an outcome or a process measure is part of that ability to identify it as unambiguous, especially given the definition that's provided in the measure's evaluation criteria and guidance for evaluation of measures for endorsement, which I presume the developers get the September 2021 version that defines the process as a systematic assessment and grading of a quantity, quality and consistency of a body of evidence that measures -- that the measured process leads to a desired health outcome.

Now, we're all in agreement that putting a person's name on the waitlist leads to the outcome, hopefully, of a transplant and a better life. And so, you know, I think it fits very nicely with that definition of "Process."

And as Zhenqiu mentioned, the way that we treat our risk adjustment, which is part of the validity part of our discussion, which we are going to be discussing and commenting on, is treated very differently for a process measure versus an outcome measure.

And so, the confusion or the lack of clarity of whether this is an outcome measure as requested by the developer or if it better fits the definition of "process" per the specifications in our technical specifications, the measure's evaluation criteria is an important determinant.

And I think that is the nexus of the discussion that we've had online regarding this issue.

The other question that relates to validity again relates to the -- what the developer reports. And, again, let me see if I can find that language.

It says: We divided the practitioner group into three tertiles based on their measure performance. And then each tertile, we computed the mortality and transplant rate.

And so, the question here is our -- the differentiation between active status, which this measure measures, and -- for the patient on that waitlist versus the next measure, '95, which apparently takes into consideration both the active patients and the inactive patients.

Is this being -- is this measure applicable for both of those -- excuse me, are the vast majority, as it would appear, of all dialysis units required to report both of these measures or just one of them?

Based on the n of about 2276, my presumption is that those -- for both of the measures, that there's a large overlap.

And so, the question is, then, the distinction between active status and general on the waitlist, what is the differentiation in terms of the performance of the Agency on these two measures.

Again, I know we're jumping to that next measure and I don't mean to compound this discussion because I think the technical information -- the technical questions that Zhenqiu raised regarding validity are very important, but the conceptual linkage between process and outcome and this determination is what is a concern.

Co-Chair Nerenz: Thanks, Gene.

Others from Subgroup 2 before we move back to U of M-KECC?

(Pause.)

Co-Chair Nerenz: No hands. Nothing in the chat. Okay. Back to our developers for -- your turn.

Dr. Shahinian: Hello. This is Vahakn Shahinian with

UM-KECC again. Thank you for the questions.

So, I'll try to tackle some of these and then I'm going to turn over for some of the more detailed statistical response to one of my colleagues and statisticians, Dr. Kevin He, and I'll -- so, I'll turn to him in a moment.

You know, ultimately our decision in terms of the risk factor adjustment, I think, was, again, intended, you know, driven by a conceptual basis for what factors may ultimately control candidacy for a patient beyond the waitlist that may be out of dialysis practitioner's control.

In order to be able to validly capture quality of dialysis practitioner performance, we want to be sure to be adjusting for factors that may lie outside their control.

And the, you know, we tackled that in several categories. Certainly patients who are sicker may not be candidates for waitlisting and, through our comorbidities, we attempted to control for that, you know, that issue.

The social risk adjustment, I realize that's been a point of concern amongst the reviewers, but transplant centers certainly take into account the availability of social support and resources and financial resources because those are fairly substantial in order to help a patient ensure a good outcome post transplant.

And so, many transplant centers do take that into account and those may be some elements that are -- may lie outside of dialysis practitioner control.

And because of that, there was a sense -- and this was certainly something for which there was consensus in our technical expert panel that adjustment was warranted to ensure that we weren't penalizing providers that were disproportionately caring for those more socially vulnerable populations.

The adjustment for transplant center characteristics was intended to, again, capture factors that might be occurring at the transplant center level or on factors having to do with organ availability that might impact -- again, that might impact waitlist candidacy in ways that were outside of dialysis practitioner control.

With respect to some of the concerns about the way the risk adjustment model was validated, you know, we're working with a -- essentially the universe of patients that this -- the measure would be directed at.

And so, it would be difficult to conceive of a completely independent set upon which to do kind of an external validation.

I'll have our statistician comment on the characteristics of our model as well as the issue around the -- how we handle the issue of the dependence within patients of the -- of the patient-month approach.

So, let me turn to Kevin and then I might turn back a little bit to process versus outcome again in a moment, but, Kevin, did you want to comment on a couple of those factors?

Dr. He: Sure. So, this is Kevin He from UM-KECC.

First, I will address the question about the correlation among patient -- the patient-month and observations.

So, the correlation among patient-months has been accounted by implementing the empirical measure.

So, this measure was originally proposed by -- analysis simultaneous hypothesis testing and then later -- extended this measure for -- either evaluation.

So, the goal of this measure is to separate the underlying unexplained variation such as the over dispersion due to the correlation among patient-months in the practitioner group that accounts from the variation that might be attributable to the quality of care.

So, when we fit the model we assume working independent correlation structure for the patient-months.

If this assumption is correct, then the tests statistics of the quality measure will follow a standard normal distribution.

We know we do have the correlation among patient-months that's why the statistics is over dispersed. Just build to the correlation.

Therefore, we did a new -- to assess the quality measure. Instead, we use the empirical --that's why people call it the empirical now measure.

And for another comment about the validation and future outcomes, again, because the aim of our risk-adjustment model is to separate the variation of the outcome that are due to the patient risk factor from variation that might be attributed to the quality of care.

So, therefore, we focus on the Gudanese fit (phonetic) of the model instead of the future prediction. So, our model updated annually. So, we focus on fit of the model assessment. And that's all. Thank you.

Dr. Shahinian: Thanks, Kevin. I think, you know, again I guess there were issues raised around process purposes.

I think that the -- again, our feeling is that the achievement of health status that reflects the ability to be a candidate for transplantation in and of itself is a beneficial health outcome.

In order to achieve that health status where you would be considered a candidate for to be a transplant, it requires optimization of your health status, you know, making sure that underlying chronic conditions are addressed.

There is a lot of effort that goes into achieving that. And so, it represents the culmination of that and represents a beneficial health outcome in and of itself.

The distinction -- also, I just wanted to make the distinction between the inactive status, sort of this measure, which is those in active status versus overall waitlisting and, again, we think conceptually that there is a difference there, you know.

Active status is a subset of waitlisting and means that you are actively able to get a transplant at that moment in time, and it requires essentially this active maintenance of health status month to month.

Getting patients waitlisted, you know, is kind of a more broad, longer-term view that also has potential chaos in terms of emotional benefit, psychological benefit of being on the waitlist.

These are things certainly that we got out of discussions with patients as part of our technical expert panel.

And so, we felt that both of these measures independently were important and we do see them as beneficial health status in their own right. Thanks.

Co-Chair Nerenz: Okay. just to check, anything else from U of M-KECC for the moment? That sounded like a closure sort of statement.

Dr. Shahinian: Yes. Yes. Thank you.

Co-Chair Nerenz: Okay. Back to, then, Subgroup 2. Any followup questions or observations based on what we just heard?

Member Lin: I actually have followup question for the developer in terms of adjusting for transplant center.

I still don't understand the rationale for that, but I'm also worried about that you adjust too much transplant center effect, you know.

Along the line about adjusting for social risk factor, you mention that transplant center incorporate the decision-making, you know, incorporate the patient social economic status into the decision-making.

So, that's why you adjust for patient social risk factor and then you adjust for the mortality ratio in the transplant rate.

And on top of that you have the, basically, transplant center effect and I'm actually worried this might explain away, or that this may potentially be attributed to the provider group, right, or look at it from a different angle. If these are so dominated by the transplant center, does it make sense to have this measure based on practitioner group, you know, information, right, because you look at the model and you can see the effect of transplant center are very strong, right? Even stronger than the Area Deprivation Index.

So, it does raise that question. So, I'd be curious to hear from the developer on that consideration on this issue.

Dr. Shahinian: So, this is Vahakn Shahinian. So, I mean, I think that we certainly see that getting a -- in terms of thinking about achieving the outcome of waitlisting, we do think that there are contributions that are important both on the dialysis practitioner side and from the transplant center side.

We're, you know, thinking about this as a -- certainly, you know, there are shared contributions and we're -- we are trying to develop, at this point, a measure that is focused on the dialysis practitioners because they play such a, you know, clearly important role.

There's, you know, conceptually when we think of the steps that are involved to achieve the ultimate outcome of waitlisting, the dialysis practitioners have roles there that are certainly outside of the transplant center control.

Getting the patient properly educated, referring them to the transplant center itself, making sure that their health status is optimized at the time they get to the transplant center, these are all things that only the transplant -- the dialysis practitioner can do, but, you know, we also, you know, have to acknowledge that there are some components of the ultimate decision to waitlist that are contributed to by factors that reside at the level of the transplant center.

So, we're trying to acknowledge the importance of the dialysis practitioner because they are crucially important and, in fact, uniquely positioned to contribute to this outcome, while, at the same time, to try to, you know, appropriately and valid the measure -- have this measure be a reflection of the quality of the dialysis practitioner who need to account for factors that may sit outside their control that can occur at the level of the transplant center. I think I'll stop there.

Co-Chair Nerenz: Okay. I'm just watching Zhenqiu for anything further. I'm not seeing a followup to the followup.

Anyone else, then, in Subgroup 2? Further questions or thoughts?

Member Lin: I have a followup question on different issue, but I just wait for -- I would be curious to hear from clinician colleague whether -- what's their view on this.

(Pause.)

Member Lin: I can go to the followup question in terms of, first, on validation. And I think over the year there are many, many risk-adjustment outcome measure came to this panel. I think most, you know, most about development in validation, right?

So, I'm not totally convinced that it's not needed for this particular measure. And other colleague can chime in on this issue.

And in terms of -- I understand Kevin refer to the -- and other, so I have one followup question.

So, for this measure score, can you specify how the measure score is intend to be used?

Are you using the score itself as point estimate or as a categorization whether they are significant different from average or no?

Dr. Shahinian: Kevin, did you want to respond to that?

Dr. He: Yes. Sure. I think that my understanding, the main purpose of for -- patient group are based on the test result. For example, we identified the extremely poor or extremely good groups based on the testing part.

Member Lin: So, basically you are using like whether they are significant different from average or, you know, better or worse, right?

Dr. He: Right. Right. And also even -- and even for the face value of the measures.

So, I guess the philosophy is similar to the equation -- the model we assume are working independent -- structure -- the same. When we do the inference we take into account the correlation. So, the correlation will effect the inference part.

Member Lin: Right. So, what I'm concerned is now if this measure score use as point estimate, that's where you not account for uncertainty to varying degree, right, and that may trip you up. That's where the lack of independence come into place more importantly.

But if you only use the categorization that might not be an issue.

Dr. He: Right. Thank you. Thank you.

By the way, for your information -- but also, you know, we're trying for the face value of the measure into standardizing the -- based on the inference and also take into account the correlation.

Member Romano: I'll address maybe Zhenqiu's question from my perspective as a clinician member of Subgroup 2.

And, again, I think the context here is that ultimately it's the Standing Committee that has the greatest expertise, in this specific domain, of what should be included in a risk-adjustment model for waitlisting.

I would definitely be on the side that I'm not sure that the decision to waitlist or to keep the patient in active status on a waitlist should be governed by the success of the local transplant center. So, it does seem a little bit odd to me to consider this in a model for waitlisting.

So, as the developers move forward, which perhaps they will here, I think this issue will need to be considered in more detail with the Standing Committee.

I mean, we certainly understand that there is some, you know, implicit way in which our conversations with patients might be influenced by, for example, the local transplant mortality rate.

I mean, it's possible that we might counsel a patient not to go on the waitlist if we thought that the transplant mortality rate at our local center was too high.

But I'm just, you know, really having trouble because the waitlisting is so much the first critical step to go anywhere down the pathway.

So, I personally, as a clinician, tend to view getting patients onto the waitlist as the essential thing that must be done and then, of course, however the transplant center behaves is sort of later on down the pathway.

But, anyway, I think that's not really an issue for our vote today, but it is a very important issue that the developer will have to confront in further adapting the measure and discussing with the Standing Committee.

Member Nuccio: Just if I can follow up, this is Gene.

I'd follow on with what I think Zhenqiu was talking about, and that is the ratio of 90-plus percent of the dialysis centers being described as As Expected and approximately five percent on either tail.

My presumption is that after you've done your analytics, then you've -- based on the tertile analysis and whatever, that you recut the distribution to describe that 90 percent as in the middle and five percent on either tail so you establish that sort of confidence level. Is that what you did?

And then, you know, mathematically and then see how the actual distribution empirically falls out.

I think this is to Kevin, but I'm not sure. Or Zhenqiu.

Dr. He: Yes. Yes. Right. Sorry, this is Kevin He from UM-KECC.

So, we use the empirical distribution of the Z score, the test that they use, and based on the empirical distribution we will flag the participants in the group into either as expected, as you mentioned, as the medial part, or the two extreme tail, the extremely better or the extremely poor.

Member Nuccio: Yeah, I think that's what you were trying to get at and I believe it's 2B.05 in your presentation.

I just encourage you to be a little bit more detailed there in the future to help us understand how the three groups were -- the three categories were differentiated.

And, you know, given the very low distribution of dialysis centers that appear to be putting people on waitlists on either active or just general status that become inactive.

Dr. He: Got it. Thank you for the comment.

Member Nuccio: And also, while I'm -- I was going to mention this in 95, the next item, but I want to thank you for including the -- here we go -- the Area Deprivation Index value.

I think Zhenqiu did mention that in his comment about both including that one and the dialysis center -- I mean, excuse me, the transplant center variable, but at least it's some recognition that social factors do make a difference in transplant success over and above, you know, getting them on the list.

Co-Chair Nerenz: Okay. Thanks, Gene. I'm watching the clock here. We're doing pretty well, actually, and I think this has been on point and focused and all good.

Anything else from our Subgroup 2 members before we move in the direction of a re-vote?

(Pause.)

Co-Chair Nerenz: I don't hear anything. I don't see visual hands up and I don't see anything in the chat.

So, Gabby and Hannah, I think it's back to you. And thanks, again, to our developer friends at UM-KECC. Very thoughtful, helpful responses. Thank you.

Ms. Kyle-Lion: Alrighty, everyone. Just give me a moment to pull up my screen. Again, as a reminder, this is just voting for Subgroup 2 on Measure 3694.

Okay. Voting is now open on Measure 3694 for validity. The options are A for high, B for moderate, C for low, or D for insufficient. And, again, we are looking for ten votes here.

(Voting.)

Ms. Kyle-Lion: Alrighty, we're --

Dr. Pickering: There's also no recusals on this measure. Sorry, Gabby. I'm just going to mention there's no recusals on this either. So, go ahead.

Ms. Kyle-Lion: Thank you, Matt.

We are at ten votes, so I'm going to go ahead and lock the poll. There were zero votes for high, six votes for moderate, four votes for low, and zero votes for insufficient.

Therefore, the measure is consensus not reached on validity still. So, I will go ahead and pass it back to Matt.

Dr. Pickering: Okay. Alright. Thank you. Thank you for the great discussion. Again, this format -- appreciate us going through this in this type of format. So, it is a CNR.

So, for the developer's knowledge, that means it goes to the Standing Committee for their assessment and vote on this validity criterion specifically since the SMP was consensus not reached.

3695 Percentage of Prevalent Patients Waitlisted

Excuse me. Okay. So, Gabby, if we can pull up the next slide, we'll go to the next measure, which is 3695. So --

Ms. Kyle-Lion: Yes. Sorry, just give me a second.

Dr. Pickering: --- similar -- yes. Sure. Similar discussions as we've had with the previous two measures. So this is also a UM-KECC measure as well.

Okay. Thanks, Gabby. Yeah, perfect. Thank you. So, this is a new measure. This measure tracks the percentage of patients in each dialysis practitioner group practice who were on the kidney or kidney-pancreas transplant waitlist.

Results are averaged across patients prevalent on the last day of each month during the reporting year.

The proposed measure is directly standardized percentage, which is adjusted for covariates such as age and other risk factors.

It is categorized as an outcome measure using claims and registry data and it's at the clinician group practice level. Does use some risk adjustments as well as a risk-adjustment model.

We won't be focusing our conversations today on reliability as the measure did pass reliability in the preliminary assessment from Subgroup 2.

The developer did conduct accountable entity level reliability testing using interunit reliability testing, and the IUR value was determined to be 0.9409.

Moving to validity, this was consensus not reached on validity. So, the developer did conduct empirical validity testing at the accountable unit level or accountable entity level, and then tested the validity of the measure by evaluating the association between dialysis practitioner group level measure performance and mortality and overall transplant rates among all patients attributed to the practitioner group.

They examined the Spearman correlation between those measures and the dialysis practitioner group level average mortality was 17.9, 18.2 and 19.2 deaths per 100 patient-years for each of the three tertiles, so T1 to T3 respectively. And then the Spearman correlation coefficient for this was -0.087 and found to be statistically significant.

The dialysis practitioner group level average transplant rate is 5.3, 3.9 and 3.1 transplants per

100 patient-years for the tertiles of T1 to T3 respectively. The Spearman correlation coefficient was 0.266, again statistically significant.

And the developer noted that higher PPPW performance correlated with higher transplant rate and the relationship with morality was also as expected by the developer and statistically significant with numerically lower mortality with higher performance on the PPPW measure although the magnitude of the association was smaller than the transplant rate.

So, you can see our lead discussant here is Eugene and second by Zhenqiu. And so, with that, we'll focus our conversations on validity and the concerns presented by the SMP.

So, Dave, I'll turn it back to you.

Co-Chair Nerenz: Alright. Thanks, Matt. It looks like we just flipped the roles of Gene and Zhenqiu for this one and I'll rely on them to keep us focused on what's same or different from the measures we've just talked about.

So, Gene, I think this is in your hands now.

Member Nuccio: Thanks, Dave.

Yeah, I was joking with a colleague that I was going to audiotape Zhenqiu talking and just play it back because we -- both Zhenqiu and others have chimed in quite clearly on what the concerns were.

These are essentially identical measures. They just represent slightly different patient populations.

The first one includes only active patients on the list -- on the waitlist and then this one here includes both active and inactive. And there are any number of reasons that a patient could be inactive, including things that are clearly outside the domain of the dialysis center, I believe. For example, the loss of insurance would often create a transplant -- have a transplant center place a patient on inactive status.

And as someone who has just recently gone through a stem cell transplant -- and I've seen what Medicare has paid for my luxury stay for 19 days at the hospital as well as all the other things -- it's quite expensive. So, without insurance it could be, you know, impossible.

Sorry about the sidebar. But the issues, I think, are the same. I think we've discussed many of them.

I did -- for anyone that's not on the -- has the chat open, I want to comment that you should read through the chat that's been going on primarily between Jennifer and Ronald and David regarding the issue of waitlisting, and whether it's an outcome and what kind of outcome is the kind of possible to perhaps measure. So, if you've not seen that information, I encourage you to do so.

I would point out that there are at least some -- there was some attempt to address social determinants of health and their risk adjustment.

And also, they did some interesting -- the developers did some interesting analyses regarding older risk -- race -- excuse me, social determinants that might be related. And I specifically call your attention to the information in Table 13 on page 41 of their document that looks at the PPPW rate, I presume risk-adjusted, but without race and with race and also by sex or gender kind of thing. It had also ethnicity.

I found those interesting tables to look at and to try to determine where the -- if you will, not including additional sociodemographic variables, how that might change the characterization --- or categorization of a dialysis center as being better than expected, as expected, or worse than expected.

And it doesn't change a lot, it does change some, and so that also suggests that some of these social determinants of health might be of additional value or at least additional investigation going forward.

Zhenqiu?

Member Lin: Well, you know, cover all. Given that we have a little bit more time, so I do have a follow-up question. It's the same for this measure and the previous measure.

One thing that puzzle me is when a developer select a comorbidity for the risk model and instead are using the outcome of being waitlisted or being active in waitlisting, they use one-year mortality. So, and they try to identify comorbidity association with one-year mortality.

So, it was a little bit puzzling. So, perhaps the developer can explain their rationale behind that.

Co-Chair Nerenz: Alright. Thanks to both of you.

Anyone else from Subgroup 2 before we move back to you then, Jack?

(Pause.)

Co-Chair Nerenz: Alright. I see a note from Jack in the chat. While we digest that, let's go ahead -- let's move back to our developers.

We have a couple of questions now being raised specifically on this measure.

Dr. Shahinian: Hi. This is Vahakn.

So, I guess the main one I heard was just the last one with respect to why our selection of comorbidities was based on the mortality, you know, the early -- kind of the early mortality and -- the one-year mortality.

And the reasoning with that is just, again, in terms of who -- what might affect transplant candidacy and that gets to, you know, the concept behind how patients are selected for transplant candidacy.

So, for patients who really derive the benefits from transplant, they need to be expected to survive a certain kind of minimum amount of time because there's an initial increase actually in mortality related to the procedure itself and it takes time for the benefits to accrue.

So, patients with conditions that place them at very high risk of early mortality are less likely to be deemed candidates, and that is conceptually why we -- our choice of comorbidities was driven based on how they related to one-year mortality.

So, that was the --- that was the decision-making around that. It has to do with clinically how that decision-making is done.

I wasn't sure that there was anything new relative to discussions we've already held. I guess one point that I would, again, reiterate with respect to, you know, the choices we made about what we're adjusting for, again, we're really acknowledging that dialysis practitioners play a crucial and unique role in helping patients get waitlisted, but, at the same time, that there are these factors there that lie outside their control and that's really our motivation in the adjustments.

So, for example, the -- you know, the waitlist mortality at the transplant center, the choice of that adjustment is based on -- that's essentially a reflection of the kinds of patients that that transplant center is willing to accept.

There's variation in how transplant centers may -- you know, their kind of tolerance for how ill patients they will accept as candidates for transplant. And the waitlist mortality is a reflection of that.

And so, if they have sicker patients, if they're willing to take on a broader range of patients, then that will be reflected in their transplant waitlist mortality.

And, again, to account for something that's occurring at the transplant center level that may be outside of the dialysis practitioner that's referring their patients to that, that's the idea behind that is to adjust for that factor that lies outside of the dialysis practitioner. I'll stop there. Thank you.

Co-Chair Nerenz: Okay. Gene, go ahead. I know Larry's got a hand up. I just want to make sure we do the Subgroup 2 folks and then we're going to get to Larry once that's done.

Gene, I think you're up next.

Member Nuccio: Okay. Just real quick for Vahakn.

The way you characterize the outcome for the providers, better than expected, as expected, and worse than expected. Currently, if I'm reading the rates correctly, you're characterizing better than expected as a group that is only presenting -- is not doing terribly well, you know. They're doing better than the rest of the crew, but they're not doing terribly well.

At what point do you expect the agencies, the developer -- for the providers, excuse me, to achieve, you know, a median value of 50 percent of their patients getting on a waitlist or is that even -- right now, I'm just sensing that better than expected is not really representing how well people are doing in objective terms.

Dr. Shahinian: I mean -- so, this is Vahakn.

I mean, I agree, you know, the way we're structuring it is more in terms of relative performance.

I think -- and part of that is because there is no -- I mean, there's no one who can define them, you know, in absolute terms what is the right, you know, waitlist rate.

Clearly not everybody is a candidate. Not everybody is going to do well or better with, you know, essentially being waitlisted versus not. So, I think it's hard to define the, you know, in absolute terms, what is the right rate.

And so what, in effect, the measure is doing is trying to identify those dialysis practice groups that are clearly doing better than the -- you know, than the average or typical dialysis practice group and conversely identify those that are clearly doing or falling short of that typical or average performance.

So, that's kind of philosophically how this is being approached because of the difficulty in clearly defining -- there are obviously measures, you know, or certain other measures in other contexts where an absolute value could reasonably be defined. I don't think that's clear here.

Member Nuccio: Sure. Thank you.

Co-Chair Nerenz: Okay. Anybody else? Subgroup 2? Zhenqiu and -- Larry, I'm not forgetting you.

Member Lin: So, I just want to ask a follow-up question, see whether it's possible. Say there are two different transplant center. One is very good, one is not as good.

So, it depends on which transplant center you work with, right? If you're not lucky, you work with the transplant center, different type of transplant center, your performance measure will be affected.

What I'm worrying about and I -- do we customize this measure to the performance level of transplant centers so that, you know, maybe we missed the opportunity to incentivize better behavior, right?

I mean, like, one group could perform really well than -- you know, but still look not as well because we account for the transplant center.

Dr. Shahinian: This is Vahakn.

So, again, I think that the issue is because there's a shared responsibility here between the transplant center and the dialysis practitioner, but, you know, ultimately, you know, we have to start, you know, somewhere.

And I think the idea here is -- initially is to hear this is a measure focused on dialysis practitioners and for exactly the point you made because that there are maybe differences in the quality of the transplant centers and there may be limitations on where dialysis practitioners may be able to send their patients, we want this to be a good reflection of the dialysis practitioner's quality.

I, you know, I think that, you know, if you take a very broad view, that there may be other measures that need to be developed at the transplant center level, but I think that you have to start somewhere when you're looking at a scenario that has shared responsibility like this one.

And here, you know, in this case, we're starting with the dialysis practitioners. They're often the ones that initiate the, you know, the process for these patients. They are the ones that are directly caring for these patients.

And so, they are the ones that are really responsible for these patients in terms of, you know, providing the best care possible.

And so, we think it makes sense to start with them, but, at the same time, during this development and in terms of the specification of the measure, we want to be sure that we're adequately capturing their quality and not penalizing them.

Co-Chair Nerenz: Okay. Larry, I think -- let's turn to you and then we'll spin back one more time before we go in the direction of re-vote. Still have a little time.

Member Glance: Thanks, David. I'll be brief.

So, I looked at the models specification on pages 28 and 29 of the document that the developers submitted.

And just to really, really briefly review, it's fairly straightforward. It's a logistic regression model, mixed effects model in which they include patient level risk factors, a facility -- or group, rather, fixed effect and then a random effect for the transplant center.

They estimate that model and then they use that model to predict for each patient in a group practice what their predicted probability of the outcome of interest is going to be conditional on the patient level risk factors, that group fixed effect, and that transplant random effect.

And then they do that for each one of the patients in a particular group and they average it together to come up with their metric.

So, when I'm looking at this -- and maybe I'm missing something and maybe others can explain this to me --- there's -- you're essentially calculating the average prediction -- predicted amount of the outcome of interest for the group.

There's no comparison of observed to predicted using the typical ways that we do this, either O to E ratio or P/E ratio or an adjusted odds ratio. It is just a prediction.

So, in essence, you could have a facility that has patients that are more appropriate for transplant and they would have a higher rate based on this model, and others that have a case mix where the patients are not very appropriate for transplant and they would have a lower predicted rate. But, again, within those individual facilities, there's no comparison of the O to E.

So, if you have patients that are not very appropriate for transplant, well, okay, so your predicted is low, but you would want to compare the O to the E as a way of quantifying the performance of that facility.

So, maybe I'm missing something here, but I would ask the folks who have spent a lot more time looking at this measure to explain this a little bit better than -- to me. Thank you.

Dr. He: Yeah, this is Kevin He from UM-KECC. (Audio interference.) So in fact, what do we use? So the example you give, people call the (unintelligible) standardization, and as you mentioned the (unintelligible) the numerator is what we observe for this group and the denominator is the respective, is a hypothetical value for all the patient within this group; however, the same (unintelligible) or no, that's what people use for the (unintelligible) standardization.

And what we use is very similar to the (unintelligible) rationalization, we call it the (unintelligible) rationalization. The numerator is that disparity of outcome across all observations.

If all the observation have the same event rate as the facility --- as the group under evaluation, then that's the numerator.

And the denominator, that's the overall, that's the overall observed number of event across all observations. That's why the interpretation is very similar to the (unintelligible) rationalization.

If this one is greater than -- I mean if this ratio is greater than one, that just means this group under evaluation has a higher event rate as compared with an average across all other groups.

And if this ratio was less than one, then that just means this group has a lower event rate. I will stop here.

Co-Chair Nerenz: Okay. Let us then cycle back. Anyone else Subgroup 2? Getting to our last call here.

(Pause.)

Co-Chair Nerenz: Alright. And I see nothing in the chat and we are nicely on time.

Hannah and Gabby, I think we're back to you.

Ms. Kyle-Lion: Alrighty, everyone. Just give me a moment to share my screen. Again, this is for Subgroup 2 voting on Measure 3695. There are no recusals on this measure.

Okay. Voting is now open for Measure 3695 on validity. Your options are A for high, B for moderate, C for low, or D for insufficient. And, again, we're looking for ten votes here.

(Voting.)

Ms. Kyle-Lion: I'm seeing nine, so just --

Dr. Pickering: Hey Gabby?

Ms. Kyle-Lion: Sorry?

Dr. Pickering: I think Jen stepped away --- did Jen step away? Is she back? Jen Perloff.

Ms. Kyle-Lion: Oh, right. I'm sorry. I totally forgot that she had stepped away.

Dr. Pickering: Yeah. So, I think we're getting nine then.

Ms. Kyle-Lion: Okay. Perfect. Then we are at the required number of votes, so I'll go ahead and lock the poll.

Voting is now closed for Measure 3695 on validity. There were zero votes for high, five votes for moderate, four votes for low, and zero votes for insufficient. Therefore, the measure is consensus not reached on validity.

I will pass it back to you, Matt.

Dr. Pickering: Okay. Thank you. So, we are a bit ahead of schedule here. So, I do want to make sure that folks are going to join on time when we reconvene.

So, right now, we have in our agenda to break for lunch. I'm just scrolling to that piece of -- okay. Sorry about that.

So, right now, we have it in our agenda to break for lunch. Originally that's at 12:50. We're about 12 minutes or so ahead of schedule.

We'll go ahead and extend the lunch a little bit longer. So, make sure to come back at 1:20 just so that, you know, other developers and such are joining during that time.

So, with that, we will reconvene back at 1:20 and the measure that's up after we come back is 3679, Home Dialysis Rate measure. And this is what we'll be starting with when we come back.

So, again, thank you all for this morning's evaluations. We will see you back at 1:20 sharp where we will reconvene and start out with 3679.

(Whereupon, the above-entitled matter went off the record at 12:39 p.m. and resumed at 1:20 p.m.)

3679 Home Dialysis Rate

Dr. Pickering: Alright. We're going to start the recording back up. Thank you, Gabby.

Alright. We're reconvening at 1:20 p.m. on the Eastern side. So, welcome back everyone. We're continuing on with our evaluation proceedings today.

We still are going to go through a few more renal measures, those measures specifically for Subgroup 2, then we'll be switching to Subgroup 1.

So, again, how this is going to work is that for measures that have both reliability and validity -- that are up for reliability and validity, discuss and re-vote, the NQF staff will present the measure and then we will have a discussion -- we will present the measure for both reliability and validity, the testing analyses that were done, and then for discussion we'll start with reliability first.

So the lead discussants will discuss any concerns or questions they have related to reliability and then we'll go to see if the developer has any responses to those questions and concerns.

We'll have a little bit of back-and-forth if needed, and then we'll go to a vote. And then we'll go on to validity and do the same for validity. So I just wanted to make note of that.

And so as a reminder as well, we kindly ask that the developer only speak when being recognized or being asked a question just to make sure that we keep some of the conversations going within the SMP participants.

Again, this is Subgroup No. 2 and with that, Hannah -- or, excuse me, Gabby, we'll go to our measure.

Okay. Great. So again, this is Subgroup 2 and it's Measure NQF No. 3679. This is Home Dialysis Rate measure.

So we're here, we're looking at both reliability and validity. Each received consensus not reached.

The lead discussants for this measure are Patrick Romano, with the secondary being Daniel Deutscher.

The developer is KCQA, which is Kidney Care Quality Alliance. This can be found in the discussion guide on page 22 as we're going through this.

So this measure, it is the percent of all dialysis patient-months in the measurement year in which the patient was dialyzing via a home dialysis modality.

And it's an outcome measure, or specifically intermediate clinical outcome measure.

The data source here is claims, electronic health data, electronic health records.

It is at the facility level of analysis and risk stratification was applied by age, gender, race, ethnicity, dual eligibility.

As far as the reliability, there was consensus not reached on this during the subgroup preliminary analysis or preliminary assessment.

The developer did conduct an accountable entity level testing, and the testing was conducted with two large dialysis organizations that could provide data as submitted to the primary data source used for this measure.

And the facility-level signal-to-noise reliability testing was conducted using the Adams beta-binomial and the mean reliability was 0.9989. So more than 90 percent of facilities has reliability greater than or equal to .99 -- or 0.99.

The smallest facilities, those less than 10 patient-months, the tenth percentile had reliability of 0.92.

The mean reliability of scores aggregated to Hospital Referral Region, or the HRR level, were 0.9943 and the minimum was 0.9435.

So, those were the reliability results and testing and I'll just summarize the validity results and testing and then circle back with our co-chair/lead discussants to talk about reliability.

For validity, it was also consensus not reached. The validity testing conducted was at the accountable entity level and the developer aggregated measures scores to obtain percent home dialysis at the HRR, the Hospital Referral Region Level, and compared this with CMS' Percent Home Dialysis Utilization by HRR, or the healthcare referral region, using the Pearson correlation coefficient. And they determined that the Pearson correlation coefficient result was 0.706 and statistically significant.

The developer also performed again at the accountable entity level, a systematic assessment of face validity of the measure score by convening a technical expert panel of nine members.

And the questions that were asked were, how likely is it that the measure scores provides a fair, accurate reflection of quality of care provided in this area? So, a total of 89.9 percent. So, eight of nine rated highly likely or likely.

There was also another question so what is the likelihood that the measure scores can be used to effectively distinguish real differences in performance between providers in the area? A total of 89.9 percent, or again eight of nine, rated highly likely or likely.

One panel member rated as unlikely and the developer provided reason for rating and a Standing Committee response.

There was some assessment of missing data, which was evaluated for one of the two large dialysis organizations representing more than two million denominator patient-months.

The missing data were rare overall and most common for discharge status, nursing home, long-term care facility residence status, as well as insurance status and those being less than five percent to .0004 percent -- .004 percent. Excuse me.

For the exclusions, when all the exclusions were applied, less than ten percent of patient-months were removed from the denominator with the estimated effect of 1.5 percentage point change in the measure score.

For the risk adjustment -- or risk stratification, rather, the developer opted to not risk-adjust the measure, but determined that stratification was more appropriate and a conceptual model was provided based on the published literature and internal analysis.

Poisson regression models were used to estimate adjusted outcomes. Age, race and dual eligibility were statistically significant, but there were small changes in the overall measure scores for justification with that approach.

Based on both the small impact on measure performance and the developer's perspective, that risk adjustment could obscure important disparities, the determination was not to risk adjust in this case, but to rather stratify.

So, again, we're going to be re-voting on reliability and validity. I will start with reliability and vote on that and then we'll move to validity.

So, Dave, I'll turn it back to you for our lead discussants, Patrick Romano and Daniel Deutscher, to present any concerns.

Co-Chair Nerenz: Okay. Thanks, Matt. And probably not much more for me to add at this point.

We had a very good, productive, tight discussion on the three measures we had before the lunch break.

While these are conceptually and clinically similar, they have some different features.

We have a different developer. So, I think we should try, as we did before lunch, to stay very tightly focused on those issues that are germane to the re-votes we do.

And I think, at that point, I can safely and comfortably turn it over to Patrick and Daniel.

Member Romano: Alright. Can everyone hear me?

Dr. Pickering: Yes, we can.

Member Romano: Okay. Excellent.

Okay. So, this is Measure 3679 to start, Home Dialysis Rate. Now, I think Matt has summarized the measure.

I want to say, at the outset, that since the measure developers are here and they've made some comments in their response, that we on the Scientific Methods Panel are very concerned about issues of consistency.

In fact, in some ways, it's our raison d'etre. It's the reason we exist to ensure a consistent approach to assessing scientific validity is used across NQF's entire portfolio-endorsed measures. So, it's something that we feel strongly about.

I think probably at least half of us are involved in the measure development or maintenance enterprise ourselves. So, we're very sensitive to NQF's fair treatment of developers.

And so, please interpret everything that I'm going to say and that my colleagues are going to say in that context.

I think with respect to the reliability issue, the members of the Scientific Methods Panel raised a variety of concerns.

I'm going to summarize them as two technical concerns and one fundamental scientific concern.

The first technical concern is that the measure is specified in an interesting way. It's specified at the dialysis facility level, except that there's a problem because dialysis facilities are often owned by parent organizations.

And those parent organizations may set up a centralized facility or program within each referral region for the purpose of managing home dialysis patients.

So, therefore, there are some facilities that are home dialysis-only facilities and there are other facilities that do no home dialysis that are within the same parent organization.

So, for this reason, the developer quite logically says, okay, well, we're going to define a different accountable entity and this is going to be essentially the parent organization within a hospital referral region.

And of course we know that there are two parent organizations that are very large that dominate this market and there are others that have a strong regional presence, but this is a very important way of specifying the measure.

And then, of course, if the dialysis facility does its own home dialysis and they're not part of a parent organization, then they're counted as a separate accountable entity.

Well, so the problem here is that they present reliability and validity statistics at the HRR level, Hospital Referral Region, and at the individual facility level, but they do not present results at the level of the entity that is being evaluated under this measure and by this policy.

So, the entity that's being evaluated is essentially a parent organization within an HRR.

So, there's a technical violation because we're given reliability metrics at a different level of accountability than the way the measure is actually specified.

The second technical violation that concerned a number of us is that the measure is proposed as a stratified measure.

The developer proposes in SP 18 stratification based on age categories, gender, race, ethnicity and dual eligibility.

This is a complicated stratification approach, but it may be supported. We'll discuss that under validity, but the reliability issue is that if the measure is proposed as a stratified measure, then we must evaluate reliability as a stratified measure because that is the way in which the measure is being presented and reported.

So, again, a technical violation because the developers have presented the reliability at the accountable -- at the level of the HRR or the dialysis facility, not a stratified measure.

Now, we could go ahead, of course, and decide to support the measure, the reliability of the measure as an overall measure. But then, of course, when we go into the validity discussion, we'll need to discuss it as an unstratified measure for the purposes of validity.

So, we need to be consistent in how we're approaching reliability and validity. So, I'm going to start out assuming that we treat this as an overall measure.

So, those are the two technical issues. I think the substantive issue that many of us choked on was that the developers use a beta-binomial approach, which is a perfectly accepted approach.

It's commonly used, for example, by the NCQA for its measures looking at the performance of specific processes for individual enrollees, but it incorporates some assumptions which do not apply here.

Specifically because the measure is specified for each month during the year of dialysis, there are up to 12 observations per month. Those observations are not independent of each other.

Now, this is not a problem with the specification of the measure. We have many measures that are specified in this way. We discussed this morning some of those measures from the UM-KECC team.

What's different here is that the developers failed to account for this issue in their analysis of reliability.

So, therefore, the reliability metric that's provided is incorrect. It is not accounting for the fact that we have this with in-person issue.

Now, there may be a variety of ways to account for that. Some of my statistician colleagues may have some ideas.

A commonly used approach is to summarize at the patient level and to treat the measure as the proportion of months during which a patient was on home dialysis.

That proportion then goes into the reliability analysis. Of course you can't use beta-binomial then. You have to use a different approach.

Another approach might be to do some kind of a three-level hierarchical model in which we account separately for the within person component versus the within center component versus the between center component.

So, I'll let my colleagues offer additional thoughts on those issues, but -- so, reliability, in essence, we have technical problems with what's presented and we have a very serious substantive problem that the approach that's used is certainly valid for other measures, however, it's not valid for this measure in the way that it's specified and our evaluation has to be measure-specific based on the specifications of the measure presented to NQF.

Co-Chair Nerenz: Thanks, Patrick.

Daniel, additional issues you want to bring up?

Member Deutscher: Not much. Thanks, Patrick, for this very thorough review and, Matt, for your review.

And I'd also like to thank the developers for their submission and also their responses to the issues raised.

I think I'm going to add maybe two points. The first one would be echoing a little bit more about the risk stratification versus adjustment issue.

And then very briefly I'll touch on the pairing -- the recommendation of the pairing of this measure with the Home Dialysis Retention measure, which will be next, I think, on our discussion list.

That may actually be more of a question to the NQF staff, but we'll get to that in a minute.

So, the only thing I'd like to add, the concerns about the recommendation to risk stratify and now testing the measure for either validity or reliability, which we are discussing now, based on stratification has been raised by Patrick. I'm not going to repeat that.

I look to the response, and what I learn from the response is although the developers are recommending that the measure should be stratified, they're also saying that stratification would not necessarily be implemented when the measure is deployed.

So, that was a little bit confusing and I would appreciate a clarification on this point from the developers.

And I wonder if, in such a case, the recommendation for stratification should simply be removed from this submission if there is already an acknowledgment that the measure might not be deployed in such a way or not used as a stratified measure.

That's currently, as one would probably expect that to a recommended measure would be the measure tested for reliability and validity as discussed before.

So, I think it would be better for the developers to make a decision. Is this a risk-adjusted measure? Yes or no. Risk stratification being a way to -- and then being consistent with that throughout the submission.

Maybe just to summarize this point, what I'm hearing is that there are important risk factors that the developers acknowledge that they cannot be easily mitigated and they were identified, Patrick named those, age, gender, race, ethnicity and dual eligibility; therefore, they should be adjusted for.

On the other hand, risk-adjustment testing from a statistical model has shown little impact on the overall measure. So, there is a recommendation to stratify.

But then, again, the measure is actually not tested as such and it is acknowledged that it's probably not going to be used as a stratified measure.

So, I'm basically trying to say that this puts me in a position of some confusion and clarifications would be helpful.

The other point that I'd like to raise very briefly is the recommendation to pair this measure with the Home Dialysis Retention and there's a very good reason that's provided, which is to avoid unintended consequences of a standalone Home Dialysis Rate measure.

So, what I'm hearing is that this measure, the Home Dialysis Rate measure, should not be used as a standalone measure, because that could have bad consequences for patients.

So, therefore, my question is -- and, again, maybe that's more of a clarification or educative question for me and maybe more addressed to the NQF staff, which I understand recommended to the developers to pair the measure or recommend pairing and not creating a composite measure.

So, what would be the case when we would prefer a composite versus to paired measure? I'd be happy to get some clarifications on this point.

This is what I'm going to raise here and this concern that I raised is relevant for both reliability and validity since risk adjustment is also under validity and there might be threats to the measure's validity. That's it for me now. Thank you.

Co-Chair Nerenz: Alright. Thanks, Daniel. As we manage our time here, we have quite a bit on the table already and this is all on the reliability side, but, still, if there are points that other members of Subgroup 2 wish to make either as additional points or as responses in some way to what Patrick and Daniel said, now is the time for that to happen.

Member Lin: This is Zhenqiu. Just one thing. If you look at the application for on page 23, so the developer did provide some stratify reliability information, right?

If you look at the table, so in one on the table and they stratified it by facility size.

So, the first row for facility was less than ten patient-month. Now, 19 facility with less than ten patient-month and yet the minimum reliability is .75 and the median is one, like this is really -- I mean, the number looks good, but actually it's crying out for attention.

So, that's where I think, you know, calling to the question of how reliability should be calculated for different situation.

In this case, one facility was very small number of patient-month, can you still, you know, trust the result based on the formula.

Co-Chair Nerenz: Thank you, Zhenqiu.

Anyone else Subgroup 2?

Member Romano: I'll just say to add -- to comment on Zhenqiu, that the most logical explanation for that finding is that those nine patient-months are in the same patient and, therefore, the same dialysis modality was used for all nine months because it's the same patient.

So, I don't know if -- I know the developer was looking for some suggestions of how to handle this issue from a statistical perspective, so feel free to add.

Co-Chair Nerenz: Alright. Good. I don't see other hands. I don't see anything in the chat.

Alright. Let us turn then to our developer, KCQA. I'll let you introduce yourselves and you can manage how you want to structure the response.

I know we've put a number of things out. We'll certainly give you enough time to speak to these issues and you're on. Your turn.

Ms. McGonigal: Okay. Thank you so much. I'm Lisa McGonigal and I'm here with my colleagues Kathy Lester and Craig Solid.

We are here today on behalf of the Kidney Care Quality Alliance and we wanted to thank you guys, first of all, for taking the time and considering our measures today. We truly appreciate the really thorough evaluation that you guys have done.

KCQA is the primary nongovernmental dialysis facility measure developer for the kidney care community.

Because of that, we are committed to eliminating healthcare and equities for individuals living with kidney disease.

Right now, there's about 45 percent of dialysis patients who are black and brown, but the same demographic group only makes up about 11 percent of home dialysis patients.

Because of this, the Biden-Harris Administration and CMS prioritize eliminating such disparities by deploying the ESRD Treatment Choices model, or ETC model. They deployed that in January of 2021.

The KCQA home dialysis measures were specifically developed to meet the needs and requirements of that ETC program. So, this gets back to the hospital referral region.

So, this is not a concept that we created ourselves. We selected the Hospital Referral Region as the level of analysis specifically for consistency with the ETC program.

This was established through federal rulemaking. Current law requires that CMS aggregate the home dialysis rate just as we have done in the measures across dialysis facilities under the same legal entity and within the same HRR.

Also, several reviewers called for the KCQA measures to be risk-adjusted. Our approach here, again, is deliberately in line with the ETC model. CMS has opted against risk-adjustment to avoid perpetuating known inequities.

And I also wanted to note that the minimum standard put forth in NQF's recent risk adjustment guidance report indicates that stratification can be an appropriate alternative to risk adjustment and we do agree and we maintain that stratification really is the most appropriate approach in these measures.

This will allow stakeholders to allow differences across sociodemographic groups and to have equity-focused interventions to better adjust the disparities.

Finally, because the ETC program will penalize providers that do not meet the benchmarks established in the program, the measure set is designed to promote steady, deliberate performance improvement over time by adjusting both sides of the home dialysis equation, both uptake and retention.

The intent of this is to provide that counterbalance to the financial pressure that facilities might feel in the program to patients who may not want or may not be clinically appropriate candidates to switch to a home modality.

We also do note that the home dialysis rate measure can stand alone, but it's our preference and our recommendation that the measures be implemented together, as you guys already discussed.

In regards to the patient-month concept, we appreciate your input on this. I've been trying to figure out what the problem was, but from what you're saying today, it sounds as if the issue is not the patient-month concept, per se, it's the approach that we took to testing that.

So, we are seeking some clarity from the SMP today on how we could perhaps better approach this as well as some of the other issues that you guys have raised.

Our intent is to put out strong scientifically valid measures and we would love to have your input so that we can get there.

We do believe that these measures are important. Our goal is to endorse them and implement them in the ETC program.

So, we're asking you guys for your insights today as well as your input on how best we can achieve these goals.

I'll turn it over to Kathy and Craig and see if you guys have anything to add.

Mr. Solid: This is Craig.

I will just add the data that we received, one of the large dialysis facilities was not comfortable giving us patient-level data, even de-identified.

So, they were only comfortable giving us facility-level data, which obviously makes it impossible to sort of correlate the intrapatient correlation.

So, that was one of the things that we had to struggle with. Perhaps we can go back and make our case again, but that's -- that's -- I just wanted to put that out there that that's one of the things that we're facing as well.

Ms. McGonigal: Thanks, Craig.

I also wanted to add in here one of the issues that -- it seems to be -- it's a little confusing.

KCQA is not a major implementer. So, we're developing this measure for use within CMS' programs.

So, we don't have access to the testing data that CMS might. So, we're a little bit limited in our approaches to testing, but beyond that when we made the statement that we can only recommend that the measure be stratified.

But since we are not implementing the measure, we cannot dictate that it be stratified. That would ultimately lie -- the decision would lie at CMS' feet.

We make a strong recommendation that it be stratified, but we cannot actually do the implementation of that ourselves.

And, Kathy, did you want to add anything?

Ms. Lester: Yeah. I mean, I hear the concerns about using the HRR level and thinking about the stratification, but I think there are two important things that we'd like you all to think about, too.

As Lisa said, we're looking to work with CMS to implement this in the ETC model and they selected the HRR level.

And I think as we were able to pull our data together from the community to test, it wasn't necessarily possible to achieve the technical issue I think that you were describing.

And so, we would hope you would look at it at this level and not let the perfect be the enemy of the good here because these are measures that are really important to the patients to make sure that this program does not result in inappropriate treatment or behavior of patient -- of facilities, that they are engaging with patients on the home dialysis side.

And the second piece is that, you know, this is an area of great inequity. I think you all know that from working with other measures.

And so, you know, one of the measures I think you see CMS not necessarily gravitating toward risk adjustment or stratification is because it could perpetuate and there is a great concern among patients that it will perpetuate the inequities that we seek today.

So, the way that we tried to get around that was the referral measure. there is no gap necessarily in care today there because we don't have the incentive structure that the ETC model will be putting forward this summer, which is to deeply penalize facilities as well as nephrologists if they do not move patients to home.

So, we need that balance and, again, I think there are lots of ways we can spend time about how to make this measure absolutely perfect and we certainly, as Lisa said, do not want to walk away from scientific rigor, but we also want to make sure we have something in these programs that do protect patients and encourage you to work with us to make that happen or to allow these measures to move forward and address some of these issues as they arise given that the programs that, you know, will be implementing those penalties begin this summer.

So, I'll turn it back to Lisa.

Ms. McGonigal: Great. Thank you.

Co-Chair Nerenz: Thank you both. I know we have a little pause here. Let me just take a second and do a quick response to some of these nice comments.

We're more than happy to work with you. And I think as our time progresses here we'll perhaps suggest some approaches. Patrick has already done that. We might expand on that. So, I think we can do that profitably. And I certainly would agree that these are important measures.

However, I think it's important to understand clearly for everyone what's under the purview of the Scientific Methods Panel and what is not.

The importance of a measure to patients, to the field, to CMS, and I'll speak bluntly, is not our problem. That is, it's very important for the Standing Committee, it's important to the CSAC I know as developers, and for all of us as citizens, and perhaps as clinicians or patients.

But the SMP reviews two dimensions of a measure, reliability and validity. And then we talk about risk adjustment as a contributor to both of those.

We cannot and do not have different standards of reliability and validity for measures that might be seen as more or less important.

So, I know we could talk, and I know your passions about the importance of these measures. But the, that issues is simply not germane to the issues that we're voting on.

Our task is to make a decision about, is the measure as presented to us reliable, basically yes or no. Is it valid, yes or no. And is the risk adjustment appropriate as a contributor to both reliability and validity?

So, as we go forward now in the next few minutes as a panel, as a subgroup, we're just going to try to clarify our thinking on those points. Because that's all we get to vote on. We do not get to vote on how important it is.

Ms. McGonigal: I just said, we appreciate your bluntness.

Co-Chair Nerenz: I try to be blunt because it saves everybody time.

Ms. McGonigal: Appreciated.

Co-Chair Nerenz: Patrick, maybe let me turn back to you now as our discussion leader, either response, further questions. And I'll just try to keep an eye on the clock for us all.

Member Romano: Sure. Well, as you suggested, you know, I completely understand the importance of measures in this space. From dialysis it's certainly something that's very important to many of our patients.

And it's also very important therefore the facility support that choice and enable that choice by the individuals they're responsible for.

So again, our focus here must be on liability. Unfortunately, you know, our goals are pretty strict that reliability has to be assessed at the level of the accountable entity.

In this case the accountable entity for most patients would be the, all dialysis facilities owned in whole or in part by the same legal entity or parent organization located in the HRR.

So, we would be looking to see that level, that unit of analysis for reliability. And then of course we'd be looking for the analysis to take into account, usually by summarizing at a patient level, the within person correlation over time.

Now, I wonder, you mentioned there are two large dialysis organizations that submitted data. Only one gave you this restriction of not having a patient identifier.

Would it be possible to address the components of reliability more effectively using data from one of the LDOs that was involved? So, that's my only sort of way out that I can offer to the developers out of their conundrum in the short term.

Ms. McGonigal: Yes. I appreciate that. And I don't want to speak for Craig. But I believe that we could. But I also would, and not at that distance.

But I would like a little additional clarity around, and perhaps, Craig, you understand it better than I do. But the, within first in liability, and so on. So, exactly what you guys are looking for in that regard.

I do think, and Craig, I'll let you speak. I do think we probably could move forward into, with the one LDO that was able to provide this data.

Mr. Solid: Do I understand, this is Craig. Yes. And do I understand though that we would, the measured entity would be the, it's not the HRR, and it's not the facility. It's the ownership entity of aggregated facilities within each individual HRR. Is that correct?

Member Romano: Yes. But every other would be reported on as part of this measure, yes.

Mr. Solid: So, if we had one LDO and they owned all the facilities it would basically be at the HRR level. Because they would be all right.

So, if we just ran the HRR level reliability, not using the beta-binomial obviously, on the individual measure using that one LDO's data, that would be sufficient.

Member Romano: Well, it would meet the technical requirements I guess. We would still have to evaluate, you know, whether the numbers are appropriate. I'm looking at Matt, and Dr. Drye, and others. But it would at least meet the technical requirements then.

Mr. Solid: Technically that would be appropriate to use the single LDO's data, patient level data, and run reliability at the HRR level? That would be equivalent to the measured entity, because they are all owned by that single LDO?

Member Romano: Right. Now of course, some might be concerned then that that measure might be overly optimistic, for example. Because if that LDO happens to have a large number of facilities within each HRR, compared with other LDOs --

Mr. Solid: Right.

Member Romano: -- then we would have trouble generalizing that estimated reliability.

Mr. Solid: Yes. So practical sort of general interpretation aside. From a technical standpoint though that would be the appropriate level to do it at. Okay. Thank you.

Co-Chair Nerenz: And, Dave Nerenz. If I could just jump in? I'm just going to draw attention back to something Patrick said in his opening set of comments.

There are some ways of dealing with this non independence problem. Like, for example, taking each patient as the unit of analysis.

But then for each patient saying that the percent of months available in a year, what percent was spent in home dialysis? That's a pretty straightforward thing. It may have some subtleties underneath. There may be other examples.

But I think we're understanding each other now correctly. The problem is not with using patient months as either the numerator or denominator definition.

The problem is, how do you deal with the non independence when you do the reliability calculations? And that was one example. There may be others.

Ms. McGonigal: Great. Okay. And, Craig, so yes. So, beyond the data binomial for that is there, if there's specific, we always look to end QS with a guidance in this regard.

And, you know, in the past this has always been, you know, has worked, getting our measures through. So, we were a bit confused that this came up. And so, is there a specific, I mean, is there something that is preferred at NQF to address this issue?

Or is this just sort of more an art than a science here that you, is up to the developer to proceed with whatever test they feel is most appropriate? Or do you guys have a firm recommendation in this regard?

Member Romano: Well, I will say that it's not an art. It is a science. And it is up to the developer to choose and to offer the measure of reliability that they feel is most appropriate for their particular measure specification. And then to justify that.

So, it's not our role to be prescriptive. However, it is our role, as Matt has described earlier, to say when a particular method is inappropriate.

So if, for example, you were to do what our Chair has suggested, which would be a perfectly appropriate thing to do, that is, to enter patient level data into the reliability analysis, then that would now be a proportion.

Therefore, you obviously couldn't use a beta binomial method, because it's no longer a binomial outcome. It's not zero one. Some patients will be .5, .75, et cetera. So, the method that you choose always has to be adapted to the structure of the data.

Dr. Pickering: I couldn't have said it any other way, Dr. Romano. Thank you. This is Matt from NQF.

Ms. McGonigal: And also, thank you, guys.

Co-Chair Nerenz: Yes. Unfortunately we find ourselves pressed for time. And we still have validity to deal with. We've not yet, we've gambled in here a little bit.

Last call then. Any comments directly related to the reliability dimension? Okay. Gabby and Hannah then, please.

Ms. Kyle-Lion: All righty. I'll go ahead and share my screen. As a reminder this is first up Group 2. We do have one recusal on this measure, Eric Weinhandl. But he is not in the subgroup. So, no concerns there when it comes to voting. All right.

Voting is now open for Measure 3679 on reliability. Your options are A for high, B for moderate, C for low, or D for insufficient. And I believe that Jen Perloff is back. So we are looking for ten votes for this measure.

Okay, we're at ten. Now let me go ahead and lock the poll. Voting is now closed on Measure 3679 for reliability. There were zero votes for high, one vote for moderate, four votes for low, and five votes for insufficient. Therefore, the measure does not pass on reliability. I will pass it back to you, Matt.

Dr. Drye: Matt, it's Elizabeth. I'm just going to jump in and -- don't mean to put our team on the spot. But one thing I'm hearing from the developers is we think that change, Patrick, was really helpful, you know, the way you framed it. We can say when something isn't scientifically sound.

I'm just wondering as a process matter, Matt, you know, we have these review cycles. But one thing that might be helpful is for the developer to get leads on ways they might approach things. Any developer, not just this developer, right.

And I'm just curious what there, we're in the cycle that we're in. But I'm just curious what we want, what we can fit, you know, if there's a way you can frame sort of how we're thinking about developer assistance?

Because it's something I know, and already focused on before I arrived last month. And we know it's really helpful. And I wanted to just convey that. Our goal is to be helpful to developers and not, you know, and as early as possible in your work.

And I'd give you a read. But we're still figuring out the process for optimizing that. So, I'm just calling out, I feel your pain as someone who developed measures for 15 years. That's what my main issue is. The less you go down that road and spend researchers and time.

So, any thoughts matter from the Chair. And we can come back to this if we have time at the end of the meeting. I don't want to slow the Committee process down.

But I just want to say I, our mind that developers could benefit from the quickest turnaround on that kind of aspect of review that we can provide.

So anyway, I just, sorry I'm putting you on the spot, Matt. I just wanted to share that this is something we're actively thinking about optimizing. Welcome people's thoughts on that.

Dr. Pickering: Yes. Thanks, Elizabeth. And that's exactly it. We, for KCQA you know that we do technical assistance calls. And even after these evaluation proceedings with the SMP or Standing Committees the recommendations that are shared by those Members to developers we definitely want to carry forward.

And, you know, work with the developers through technical assistance calls on those recommendations, whether it be something that, how you frame things within the measure submission to address those concerns, or even with related to some of the testing.

So, we do review preliminary testing results at NQF prior to measure submission to provide some input on what may be expected, or what may be a nice to have versus a need to have within those testing results and information shared. And that includes some of the methods and approaches that have been, that you're thinking about.

So, even after the call today there was a series of recommendations shared by our SMP Members on approaches to doing the reliability testing, and accommodating for that non independence with patient month, or accounting for.

This is something we can work with the developer on even after this meeting to ensure that, you know, the approaches are sound. And, you know, gearing up for the evaluation of the measure when it comes back.

Ms. McGonigal: Wonderful. Thank you. Yes. It, we appreciate the sentiment too. It is difficult on this end. So, we do appreciate that you guys are thinking about these things. And appreciate anything, any guidance that we can get in this regard. So, much, much appreciated.

Dr. Pickering: Great. We'll be documenting those recommendations as well, and so that we can share those with our KCQA colleagues. So, thank you.

I want to kind of circle back now to going to validity. So, I presented the validity testing and the results there. I know that we sort of started touching on some of the concerns.

But I will now focus our attention to validity for this measure, since it was a CNR. So, Dave, I'll turn it back to you for our lead discussion, to discuss the concerns or questions they may have for validity testing.

Co-Chair Nerenz: Okay. Certainly. And I'll just turn right back to Patrick and Daniel. And I think if we have already picked up some of those things in our earlier discussion we can skip over that. But I'll leave it in your hands about how we get as quickly as possible to a validity decision.

Member Romano: Okay. Thank you. So, now things get a little more complicated. Because it's not a straightforward question of appropriate and inappropriate.

So, for validity testing there are two sets of data that was presented. There is empirical validity using correlation with CMS data on home dialysis utilization rates.

And there's base validity measure. Base validity based on expert panel process. So, and then we're going to talk about risk adjustment.

So, first of all I think that we should all be able to agree on the Methods Panel that the empirical validity assessment is not helpful for this discussion.

And I say that because, two reasons, one is that the empirical validity test was done at the HRR level, not the LDO HRR, but the entire HRR level. So, it's not done at the level of the accountable entity.

Second, the empirical validity testing is essentially correlating the measure with a previous version of itself. So, it's correlating this measure with a prior year CMS data on home dialysis utilization, finding a reasonably strong correlation, as we would expect.

But that correlation is not helpful to understanding the validity of the measure conceptually, because it doesn't illuminate the structure process outcome relationship, or the relationship with subsequent patient outcomes of the choice of home dialysis.

So, we can't use the evidence on empirical validity. However, this is a new measure, so that's perfectly fine. So, we can rely on face validity.

So, face validity here was measured through an expert panel process. And there were two concerns that were raised by the panel about this process.

One is, that in general we prefer an arm's length between the developer and the expert panel. So, for

CMS measures, for AHRQ measures, for example, there's a very open and transparent process by which members of the expert panel are solicited and reviewed, and then selected.

And then of course, the minutes of those meetings or summary of those meetings become open transparent documents that anyone can review. So, there is this sort of arm's length relationship.

In this case, the KCQA convened its own expert panel. And so there's some debate about whether there's a sufficient arm's length relationship.

I think they defend the way in which the expert panel was conducted in their response. And I'm prepared to accept that. This is a common situation that measure developers confront.

What I'm a little bit more surprised by, to be honest, is something the developers didn't respond to. That is, the expert panel did not include any patients, any dialysis patients or caregivers of dialysis patients.

And this is of course really important for patient centered measures that really reflect patient choice, like home dialysis. There are of course patient representatives who were involved in other Steering Committees within the NCQA, the KCQA, I'm sorry, the KCQA Steering Committee.

So, I don't mean to say that patient and caregiver voices are not heard within KCQA. They just weren't heard apparently on this particular expert panel.

Therefore, since we're asked to evaluate the vote of this expert panel, which was a near unanimous eight of nine vote, that is perhaps a deficiency in the process.

But nonetheless, I think that the case validity information is well presented by the developers. And so, I encourage my panel members to consider the information that they presented.

Now finally, the third set of validity issues that the panel focused on was about risk adjustment. And in particular the fact is the measure developers here went through an extensive process of reviewing the literature, and basically identifying the potential value of risk factors. And even estimating a risk adjustment model.

But then at the end of this they basically say, well, you know, for a variety of technical and strategic reasons they gave up on risk adjustment, and fell back on stratification.

So, with respect to stratification of course, if the measure is actually being proposed as a stratified measure then of course our validity analysis has to focus on the measure as a stratified measure.

Some of these stratifications may be more arguable than others. I think there are also a variety of other clinical factors that aren't included in this stratification.

It may have a strong effect on the appropriateness of home dialysis for an individual patient. And so, some of those issues are explored in the developer's document.

But so, I think that's sort of my summary of the concerns. Daniel, do you want to add to that?

Member Deutscher: Thanks, Patrick. No, I really don't have anything important that I would go. I'd mainly be repeating myself. So, no, not at this time. Thank you.

Co-Chair Nerenz: All right. Thank you both. Anyone else, Subgroup 2, additional comments, questions? Okay then. Perhaps I'll turn to Lisa. Any specific responses to Patrick? He did suggest that some adequate responses had already been given in writing. But this is your chance to embellish on that if you wish.

Ms. McGonigal: Okay, yes. No. And thank you. Thank you, Al. Thank you, Patrick. Just to clarify. So, the empiric validity, you know, we submitted, we thought that it may just bolster the face validity that we submitted.

So, the issue around the face validity could be addressed. The main problem that you guys had with this is that there wasn't a patient that responded. I just wanted to be clear on that. Correct?

Member Romano: That's the remaining problem, after -- But I'm not sure. Because I think a couple of my colleagues on the panel felt more strongly about the lack of an arm's length relationship. So, they should speak as well potentially.

Ms. McGonigal: Okay. And I'll, actually, I'll let Kathy speak to that, if you'd like, Kathy?

Ms. Lester: Sure. I think the challenge any organization, and Patrick, you said this well already, that is, within the community developing the measure to be arm's length, you know, the KCQA with maybe one or two exceptions has the vast majority of the kidney care community involved.

So, the only way to be truly arm's length is either to go outside of the kidney care community, you know, or try a path on non expert.

So, I would encourage more guidance on that. And to think about a way that, you know, you can leverage the expertise you have around the table that you're working with. I would assume the AMA and others have similar problems to this over time.

And so, if the SMP and NQF staff in particular can be helpful on how to address that without losing the benefit of a measure developer in that area of expertise, I think that would be particularly helpful.

Ms. McGonigal: Thank you.

Member Romano: And could you address the reason for not including patient or caregiver representatives in the expert panel?

Ms. McGonigal: Yes. We actually sent out requests for responses for that. And then we were left with the responses that we got. We did have several patients on the workgroups, the related workgroups and the Steering Committee.

But when we sent out the call for face validity, you know, we had the nine responses. And there was not a patient on that. So, it was just sort of a, it just was the way that the responses came in.

I also wanted to --

Ms. Lester: Lisa, can I jump in here? It means that --

Ms. McGonigal: Sure, sure.

Ms. Lester: -- the patients that we were working with did not choose to participate. We did solicit their input. So, it would be similar to someone missing a meeting or two here, but still being included in the process.

Ms. McGonigal: Correct. Thank you. Thanks, Kathy, for that clarification. And then, the other thing I just wanted to bring up, the way that we approached the conceptual model and the risk assessment.

We, this was taken directly from, I believe it was in August NQF released a report on how to adjust a risk, how to address risk adjustment, particularly social and functional.

We borrowed almost verbatim from that --

(Loss of audio)

Ms. McGonigal: -- how we were really using and adhering very closely to NQF guidance in this regard. And I'll let Craig speak to the fact that the reason we opted against actual adjustment is, when we applied the adjusters to the measures it did not impact measure scores appreciably. So, we determined that the adjustment was not required. We didn't just abandon it. We determined it was not needed.

However, we did feel that stratification was needed, because there was the difference across the scores. And we know that there are disparities in this area. So, we thought that it is important to stratify. Craig, did you want to speak to that?

Mr. Solid: Well, I think you stated it. Again, this gets back to we didn't have patient level data from one of the LDOs. And so, we had facility level within each risk strata.

So, we had to run multiple models so nothing was adjusted for everything at one. And we just felt that with, it didn't move the performance that much. And so, we opted for stratification.

But again, I think, as Patrick mentioned, some of the issue is that it wasn't at the proper measured entity as well. So, I think that was, you know, we can address that moving forward too.

Co-Chair Nerenz: Okay. Thank you. I was feeling --

(Simultaneous speaking)

Dr. Pickering: Sorry, Dave. This is Matt. I just wanted to chime in just briefly here. So, just for the panel members and developers on the call. So, face validity as we've stated is a acceptable form of validity testing for the accountable entity level for new measures.

Our criteria don't specify that you must include patients, right. We just state that they must be identified experts. So, there may be concerns that the patient community is not involved, depending on the measure and the scope of the measure.

But I will say that our criteria doesn't get that specific as, you must include this stakeholder group, this stakeholder group, this stakeholder group. It does say that it should be by identified experts. So, I'm throwing that out there.

And I'll also say that, again, this is an accountable entity level testing that was conducted empirically. And then also face validity. So, since there isn't empirical assessment it does sort of start out as an opportunity to rate this as a high.

So, with the face validity assessment, you know, or with the understanding and concerns of the empirical validity, the face validity may be an acceptable form for lowering that rating from a high to a moderate, for example.

I just wanted to throw that out there as we start to get closer to a vote around face validity. That it's not proscriptive as like you must include patients in our criteria. So, you have to take those considerations into account, as well as, since there's these two forms, empirical and face validity.

If the empirical isn't looking as good, then the face validity is something that you can consider passing the measure on, if that is, if you feel there are, the concerns have been addressed. I just wanted to state that, Dave. Thank you.

Co-Chair Nerenz: Well thanks, Matt. That was a couple of points I was thinking to make as well. So, that's good. Before closing this out, additional last opportunity. I see Patrick's hand up. And then I'll go to anyone else in Group 2 that has another comment or question. Patrick next.

Member Romano: Yes. Just a couple of comments, again with respect to focusing on the risk adjustment question. So, I think that in general, of course, we find it easier to interpret risk adjustment that's done at the patient level, accounting in this case for hierarchical structure of the data, the fact that you have up to 12 months observations for each patient.

So, we understand that you were unable to do that. And you have tried to account for over dispersion in the data using a quasi-Poisson regression model that --

Anyway, that's something that you may wish to explore a little bit further if you can get the patient level data. You can construct a hierarchical model.

But I'm wondering if you could also assess your conceptual model in Figure 1. You know, properly identify some clinical variables that are very important as well, you know, related to whether home dialysis would be an optimal choice for an individual patient.

And in your actual analysis you focused on the social factors and demographic factors, age, gender, race, dual eligibility. But I'm wondering if you could comment on the clinical factors?

Obviously, you know, you mention if somebody's demented, if somebody's blind. Depending on, you know, somebody's level of obesity, for example. That may affect the appropriateness of home dialysis.

So, could you address why you wouldn't want to include some clinical factors if you were able to get patient level data for the analysis? Because I would theorize that your model would be much stronger then.

And in fact, you might find that it did make a big difference in terms of evaluating the accountable entities if you were able to account not just for demographic factors, but also for clinical factors.

Ms. McGonigal: Thank you. That's most appreciated. This again comes down to an issue of what is available to us as a small major developer without the access to the levels of data that a larger developer like CMS does.

We could not easily access the clinical variables that you're speaking of to assess them. And so, we were, we truly were unable to get that data.

We did try. But we were unable to get them from our participating organization. So this is something again where there is this bit of a disconnect from being a developer, but not implementer.

And so, if it came time to implement this measure, and it was determined that CMS thought that these clinical variables should go in, that could be reassessed at that time. But we were unable to access them. And, Kathy, and Craig, anything to add there?

Mr. Solid: Nothing from me.

Ms. Lester: I would just add that I think, you know, as we look at some of these factors, again, when we looked at this measure we were looking about closing significant gaps, not minor gaps but significant gaps in care.

And while things like dementia obviously will have a role, we're also not looking at 100 percent being your benchmark that the measure will be evaluated again.

Whereas, obesity is going to be something that clinical some physicians will say it is appropriate to have a patient on home dialysis, even if they are obese.

And as we worked with our expert groups they did not feel that these other clinical indicators were necessary at that time.

And again, I think it does go to the heart of the fact that there is such a great health disparity, particularly based on communities of color that we just didn't go as far down that path as we might have, in addition to Lisa's concerns about being able to access that data, given that we are more of a community based rather than a Federal Government based measure developer.

Co-Chair Nerenz: Okay. Thanks. Others of Subgroup 2? Going once, going twice. All right. Gabby and Hannah. Gabby or Hannah.

Ms. Kyle-Lion: I'll go ahead and share my screen. Again, this is a reminder this is for Subgroup 2. And Eric Weinhandl is still recused from this measure. But like I said, he's not in the Subgroup, so that should not impact voting.

All right. Voting is now open for Measure 3679 on validity. The options are A for high, B for moderate, C for low, or D for insufficient. And I believe we're looking for ten votes here.

All right. We're at ten. So, I'm going to go ahead and close the poll. Voting is now closed on Measure 3679 for validity. There are zero votes for high, five votes for moderate, three votes for low, and three votes for insufficient. Therefore, the measure is consensus not reached on validity. I will pass it back to you, Matt.

## 3697 Home Dialysis Retention

Dr. Pickering: Thank you so much. Okay. So, moving on to the next measure. Gabby, will you pull the slide up? This is also a KCQA measure. It's 3697 Home Dialysis Retention Measure.

So, with this there's going to be probably some similar concerns noted for this measure, as it was for the previous measure. So, I'll leave it up to Dave and our lead discussants to see if we can just see if there's any additional conversation needed, or if anyone wants to change their votes based on the conversation we had for 3679. But I will at least introduce the measure.

So, this is a new measure. We, as you can see listed, it's no path to reliability. There is a CNR for validity. Lead discussants here are Patrick Romano as well as Daniel Deutscher. And you can find this on Page 25 of the discussion guide.

This is a new measure. It's the percent of all new home dialysis patients in the measurement year for whom greater than or equal to three consecutive months of home dialysis was achieved. The new patients are defined as those who started a home dialysis modality during the measurement year.

It is a intermediate outcome measure using claims, EHR data, registry data, and ESRD Quality Reporting Systems, and that legacy CROWNWeb Clinical Data Repository.

It's, the level of analysis is defined as the facility level. And they applied risk stratification for age, gender, race, ethnicity to eligibility.

For the ratings for liability it was not passed into the low rating. The reliability testing was conducted at the accountable entity's level. And testing was conducted with those two large dialysis organizations as previous, with the previous measure.

And a facility signal-to-noise liability testing was conducted using Adams data binomial. And the mean reliability was 0.5241. So, 50 percent of facilities had a reliability score of one. And then, by sample size the mean reliability varies from .41 to .66. And the mean reliability scores aggregated to HRR level was .03787.

For validity it was consensus not reached for validity. And the validity testing here was the Systematic Assessment of Face Validity of the Measure Score. So, they convened an expert panel of nine members who were given the specifications, measure scores, and performance distribution, and asked very similar questions as the previous measure.

So, the first being, how likely is it that that measure score provides a fair and accurate --

(Loss of audio)

Dr. Pickering: -- this area. The total is 77.77 percent, or seven of nine rated highly likely or likely.

The second question. What is the likelihood that the measure score can be used to effectively distinguish real differences in performance between providers in this area? 77.77 percent, or seven of nine also rated highly likely or likely.

And two panel members rated as neither likely nor unlikely. And the developer notes that the paired measure set was rated as highly likely or likely by eight of nine panel members.

They did an assessment of missing data as well, noting that missing data were rare. And then when applying all the exclusions, approximately five percent of patients were removed from the denominator, with an estimated effect of 2.8 percentage point change in the measure score.

The developer did not conduct independent risk adjustment for the measure, stating that the home dialysis retention measure denominator is built from our Home Dialysis Rate Measure numerator. And as such they did not separate the risk, adjustment analysis for the retention measure.

So with that, again, we'll start with reliability. And then we'll move to validity. So, Dave, I'll turn it to you. And then we can begin the discussions for concerns related to reliability.

Co-Chair Nerenz: You know, just the person who passes the baton. I think all I'd ask back to Patrick and Daniel is, let's keep us focused on what's unique or new to his measure, as opposed to what we just discussed.

Obviously they're very similar. They're designed to be used as a pair. So, as we move along let's just try to focus now on what is different or unique with this one.

Member Romano: Excellent. Thank you. And yes. So, I think that we can move quickly to vote on reliability. I just want to raise two points.

So, in this case if you look at Page 25 of the measure submission you can see that the majority of facilities, 1,646 out of 2, 581 facilities have a denominator less than ten.

These facilities, which presumably have perhaps only one patient on home dialysis, have a median score of one, median reliability score of one.

So, you know, obviously this cannot be true. So, it highlights, as you look down the table you can see that the median reliability goes from one for the smallest facilities down to 0.148 for small facilities, 0.27, or perhaps small but mid size up to 0.54.

So, the point here is that not only do we have a problem of the wrong method. But we have a problem that I don't see how you're going to fix this. I don't see how you're going to get reliability to the range that we're looking for of .6 or higher. It just doesn't seem feasible.

Even with this measure, which is a highly optimistic measure, you're at .524. So, that I think what motivates some of the comments that you heard, is why not think of this as a composite measure?

The idea of pairing is critically important, and it conceptually makes sense. But again, I think from the methodologic perspective we really don't see any way that you can get to a reliable measure for accountable entity level reporting, given the size of these denominators at the facility level.

Now, it may be that when you roll those facilities up into all the facilities in the same LDO, in the same HRR you might be able to get to an adequate level of reliability. So, that would be, you know, input and challenge for you.

The other thing I just want to point is that there's something a little odd about the specification I'd like the measures, the developer to explain.

Usually in measures we apply exclusions at the denominator level. And then those exclusions roll through to the numerator. On Page 17 the flow chart shows that there are some exclusions that are applied to the denominator separately from the numerator.

In other words, first the measure is evaluated according to the numerator or denominator. And then the denominator has exclusions applied to identify patients who were discharged prior to achieving three consecutive months for specific reasons.

I would expect that those same patients who were discharged from the treating facility before three months of dialysis for specific reason, those same patients would need to be excluded from the numerator as well, since we usually think of the numerator as being a subset of the denominator. So, perhaps the developers could address that technical issue.

Ms. McGonigal: So, yes. Let me do that, and then I'll pass the floor over to Kathy. So, the technical issue that you described, Patrick.

The way this is structured is, we applied the exclusions only to those patients. We look for the exclusions only in those patients who didn't reach the three consecutive months.

Anyone who did achieve three consecutive months is already going to be in the numerator. So, there's no need to apply this to the numerator as well, if that makes sense. I think I understood your question correctly.

Member Romano: Okay. Yes. All right. So that, I'll think about that for a minute. But that should resolve the problem.

Ms. McGonigal: Okay.

Member Romano: Usually in the flow chart we show it as the numerator being a subset of the denominator. But, okay.

Ms. McGonigal: Yes. Yes, I know. We did this just a little bit differently. And I want to just pass over to Kathy for your first question as well.

Ms. Lester: Yes. And I think this is just one of those, the challenges of being sort of locked in a box. And I understand and hear that you all feel that you have those constraints.

But this is the example of something that is just so critically important to patients, right. They don't want to be put on home dialysis if that is not medically appropriate.

And the way the incentives and the ETC model are, they have raised this concern directly, and through our colleagues at Kidney Care Partners, that the incentive structure will force patients to receive treatment that is not appropriate.

We saw this happen with vascular access measures. So, they're not, you know, dreaming. They are legitimately concerned. And I guess our question is, we know we don't have this problem today, right. We would have not had this gap in treatment if we did.

But when you look at, you know, when the measure comes out it is next to impossible I think to get to the reliability level that you all need to see when we're anticipating that future problem.

And so, you know, that's where, we're not asking for less rigorous standards, but a standard that maybe accounts for this reality. So, we are proactive, and we don't force patients to suffer in the interim while we wait for the problem to occur.

Ms. McGonigal: I also wanted to, about the composite versus the paired. We did discuss this with NQF staff.

It was agreed, because we have two separate measure scores, that this was more of a paired construct.

But it sounds like you guys believe rolling it up into a two composite might be a more appropriate approach. I just wanted to make sure I was understanding you correctly.

Member Deutscher: I'll just quickly comment on that. And it may be just my view, or my misunderstanding of how NQF advises about this.

But my, I would say my intuitive thought is that if you're thinking that one measure should really be used only with the other, this is your recommendation, then why not a composite?

So, and my question again to NQF is, what is the difference, what is the practical difference between two paired measures and two measure within a composite?

Dr. Pickering: Yes. So --

Ms. McGonigal: Is that --

Dr. Pickering: One -- Oh, I'm sorry, Lisa. Were you going to --

Ms. McGonigal: I was just going to ask if that was a question to NQF. I thought so.

Dr. Pickering: Yes.

Member Deutscher: Yes. Yes.

Dr. Pickering: Yes. This is Matt. So, one of the biggest features of the composite is more of the all or none type of assessment, where two measures who can stand alone independently because they have a numerator denominator specifications, are combined into a composite measure because both of them need to be, both of them need to happen in order for a score to be aggregated.

So, for example, you have a diabetes optimal care, you know. Each one of the components in that must happen in order to achieve the composite score. So, it's an all or none type of assessment.

Whereas, you know, the non composites are usually like the ending or not. So, it's or type of relationship. Neither this or this. Whereas, the and you have to meet all of these. That's one of the underlying features of that.

In addition, it's really the value of these components together to reach some sort of positive healthcare quality improvement, right.'

So, that's part of the construct relationship. And the rationale for the construct is that these two components together lead to some beneficial outcome.

So, you can combine measures and create an and type of combination for a composite. But ultimately if the measures should be together, you know, the overall composite reflects overall better care as the two individually, then it should be a composite.

So, it's those types of considerations is what NQF in our criteria look at, this all or none versus any or none. And it just, again, the overall rationale and specification that measures or components be combined to provide an overall assessment of quality is needed and justified.

Member Romano: I just want to suggest again that this is purely my way of friendly suggestion. Because I, you know, I feel your pain, so to speak.

But in some cases as measure developers when we face this conundrum we try to put the measures together into a single measure. So, it's not formally a composite measure according to NQF criteria. But it is a single measure.

So, for example, let's play this out for a second. So, you're counting in the previous measure, you're counting all the months during which a patient stays on home dialysis over that period of up to 12 months of follow-up.

But let's say hypothetically that you don't want to reward dialysis facilities for throwing people on home dialysis who are destined to fail. And so, they only stay on dialysis, home dialysis for a month or two, right.

So, you could say from the patient centered perspective that if the patient stays on home dialysis for less than three months you zero out those months when they were on home dialysis.

In other words, you don't give the facility credit for those months. Because the facility was trying to game its numbers by throwing everybody onto home dialysis. And they were inflating the numerator for the previous measure.

So, you could take the previous measure, and you could redesign the numerator of that measure to account for this potential unintended consequence.

If you're concerned, again, it has to be viewed in the context of, you know, balancing risks and benefits, and costs. But it would be possible to redesign the first measure to basically not give facilities credit for putting patients onto home dialysis when that decision didn't last three months.

Ms. McGonigal: Thank you.

Co-Chair Nerenz: Okay. I'm detecting a pause in the flow here. Oh, Gene, I see a --

Member Nuccio: Yes. Real quick. And it's kind of a follow-up. This is Gene Nuccio. A follow-up to Patrick's point.

A difference that I noted between the previous one, the previous measure, the 3679 and this one, 3697, is that this specifically refers to all new home dialysis patients. Whereas, the other one does not.

And I guess my question as it relates to this particular measure is, what happens when a patient is no longer new? Are they dropping off the measure for this case? They would not be dropping off the measure in the previous metric. And so, I think you need to figure out which way you want to go.

Another point about the new patients dropping off is that a dialysis center could improve its margin by simply ignoring all those patients who they missed in the initial go round, and start something new. And suddenly appear to be much better than what they've done in the past.

So, just, again, a clarification for the measure developers. You need to decide whether you're tracking new patients or all patients.

Ms. McGonigal: Well exactly. And I appreciate that. For the second major, the retention major, it was specifically to look at those who are just starting on dialysis to see if the facilities are doing a good job of preparing and educating them, so that they at least get to that three month mark.

After a patient is a new patient they will be captured in that first measure. So, you know, new patients are captured there as well. But it would also capture prevalent patients in subsequent years.

So, as the rates start coming down that will also be an indicator that you're losing your patients on, who are not being retained beyond a year. So again, it's, we're just trying to figure out exactly how we balance this. It's very difficult.

(Simultaneous speaking)

Mr. Solid: Lisa, I'll also say that there was a lot of concern, and there's a lot of effort trying to put into giving facilities credit for what they are doing.

So, if you're measuring in a year, and they have patients who have been in home dialyzing, and they start in January and then they drop off, you don't necessarily want to penalize the dialysis facility for things like that.

There was a lot of concern from facilities about that. So we, some of this was in an attempt to sort of give credit where credit was due, and not inappropriately penalize just because of timeframes, or windows that we were looking at.

Co-Chair Nerenz: All right. Let's let me try to push this along. Thinking also of our four colleagues in Subgroup 1 who are going to get squeezed if we don't move this along.

I think we've addressed the issues, at least I was hearing on issue of reliability here. Is there anything last that someone feels is important on that question?

Member Deutscher: Just a very quick comment, you know, David, about reliability. Just as a technical suggestion to the developers.

Within the Adams tutorial there's, I think towards the end of it, there's also a suggested method for continuous outcomes, not on the binary outcomes.

So, if you decide to take that forward, and maybe use the outcome on a patient level as a ratio, as Patrick suggested before, you could look at that and see if it is appropriate. Just a small technical note.

Co-Chair Nerenz: That's very good. Thank you. All right. let's see if perhaps we can call the question on reliability. Anyone, last things? All right. Gabby.

Ms. Kyle-Lion: All right, everyone. I will go ahead and share my screen. Again this is for Subgroup 2, and Eric Weinhandl is recused on these measures as well. But he is not in the subgroup. So that should not impact voting.

Voting is now open for Measure 3697 on reliability.

Your options are A for high, B for moderate, C for low, or D for insufficient. And we're looking for ten votes here.

We're at nine, just waiting on one more. We're at ten. I will go ahead and close the votes. Right. Voting is now closed on Measure 3697 on validity.

There were zero votes for high, zero votes for moderate, nine votes for low, and one vote for insufficient. Therefore, the measure does not pass on reliability. I will pass it back to you, Matt and Dave.

Dr. Pickering: Thanks. And I'll just turn it back to Dave for any comments related to validity. Because we still have to vote on that. Again, if there's no, nothing in addition to what's been discussed in the previous measure we may be able to make up a little time. But, Dave, I'll turn it to you and our lead discussant.

Co-Chair Nerenz: Yes. And I just go right to Patrick and Daniel. Have we discussed the validity issues already in the context of the prior one? Or do we have any new things?

Member Romano: I think we've discussed the issues. In this case I think the team had a little more difficulty with risk adjustment, because of the small numbers problem.

But again, there are currently demographic factors that are highly associated with retention on home dialysis. And so, we can defer ultimately to the Standing Committee to decide about the benefits of stratification versus adjustment of those factors.

Member Deutscher: Nothing to add from me, yes.

Co-Chair Nerenz: All right. Anyone else, Subgroup 2? Or have we covered this ground to your satisfaction? All right. I think then we get to Gabby again.

Ms. Kyle-Lion: All right. I will go ahead and share my screen. Again, as a reminder, Eric Weinhandl is recused on this measure. Voting is now open for Measure 3697 on validity.

Your options are A for moderate, B for well, or C for insufficient. And again, we are looking for ten votes here.

Co-Chair Nerenz: I'm sorry. Dave here. Just clarification, since we skipped so quickly. Was there also face validity from the expert panel on this one, as there was for the home dialysis measure?

Dr. Pickering: Yes, there was face validity assessment on this measure as well as the home dialysis.

Co-Chair Nerenz: Yes. Okay. Thought so. Just wanted to be sure.

Member Romano: It was the same issue.

Co-Chair Nerenz: Yes, yes. Got it.

Ms. Kyle-Lion: And we're at nine votes. So, I think we're just waiting for one more. Let's give it another second. I see we're still at nine. I'll give it one more second, and then we can move forward.

Okay, we're holding at nine. I'll go ahead and close the poll, because we still have a quorum for it. So, voting is now closed on Measure 3697 for validity.

There were two votes for moderate, five votes for well, and two votes for insufficient. Therefore, the measure does not pass voting. I will pass it back to you, Matt. Thanks, everyone.

Dr. Pickering: Okay. Thank you so much. So, thank you so much to Subgroup 2. I very much appreciate the conversation. As well as thank you to our developers on the call, KCQA, as well as you and Kathy. So, thank you very much.

## Measure Evaluation, Subgroup 1, Renal

We will now move to Subgroup 1, in which we have a series of measures to close out our day. So, thanks Gabby and Hannah.

## 1460 Bloodstream Infection in Hemodialysis Outpatients (CDC) Perinatal and Women's Health

The first measure that I think is 1460. Okay. There we go. Perfect. So, the first measure up is 1460. So, you can see here we have reliability being no pass and validity being consensus not reached.

We'll follow the same format as we have been working with. The developer of this is the Center for Disease Control and Prevention, so the CDC. Are colleagues from the CDC on the call?

Ms. Benin: Yes, we're still here. Thanks, Matt.

Dr. Pickering: Okay. Thank you so much for your patience.

Ms. Benin: Yes.

Dr. Pickering: Yes. Thank you. So, I will summarize both the results for reliability and validity testing. And then we'll turn it over to Christie, who's our Co-Chair facilitator of this, of these measures.

And then turn it over to Jeff, who's our lead, or excuse me, Terri for our lead discussant on reliability. And then we'll vote on reliability.

And then we'll go to validity. Discussant concerned with validity was Jeff and others in the subgroup. And then we'll vote on validity.

All right. So, for 1460 this is a maintenance measure bloodstream infection and hemodialysis outpatient. And the description for this measure is that this is an annual measure which provides the standardized infection ratio of bloodstream infections, or BSI, among patients receiving maintenance hemodialysis at outpatient hemodialysis facilities.

BSIs are defined as positive blood cultures of hemodialysis patients which are reported monthly by participating facilities. The SIR, that's standardized infection ratio, is reported for a yearly period, like a calendar year. And it's calculated by dividing the number of observed BSIs into the number of predicted BSIs during the year.

And then an outcome measure, the data source are paper medical records. And the level of analysis is the population level at the regional or state level. They also have some risk stratification, risk categories by vascular access type.

And for reliability the measure was not passed, receiving a low rating. So, the developer's validity testing has served as a demonstration of data element testing for reliability.

So, keeping that in mind here what we'll be doing is going through validity, and talking about the data element level testing for validity. But as we're voting that's, those are the data taking into account when voting on reliability.

So, for validity at the data element level the developer has conducted the patient encounter level testing using inter-abstractor reliability.

The developer calculated a percent of BSI under reporting over multiple time periods for a national sample as of, starting back in 2015, and all going to 2020. They also reported state level data for Tennessee in 2014, Georgia in 2015, and Colorado in 2017.

The developer also validated vascular access types, so fistula, graft, tunneled central line, and others. And the developer calculated BSI under reporting of 33.3 percent for 2015 data, 16.7 percent in 2016, 52.2 percent in 2017. I'll just summarize more recent here that in 2020 was 33.9 percent, 2020.

At the state level under reporting of BSI was 58 percent in 2014 and 29 percent in 2015, and then 22 percent in 2017.

So, concordance with vascular access type reported was 80 percent for fistula, 86.3 percent for graft, 93.3 percent for tunneled central line, 96.5 percent non-tunneled central line, and 98 percent over access, other access type.

The pooled sensitivity was high for fistula, graft, and tunneled central line, all rated at 80 percent, ranging from 81.2 to 91.6. And the developer notes overall improvement in national BSI under reporting over time, with the exception granted for the first six months of 2020 due to COVID-19.

The developer notes that state level BSI under reporting showed improvement over time. And the developer notes that all access types had at least 80 percent match, demonstrating high concordance.

So again, we'll use the data element validity testing in our discussion for reliability. And then we'll vote on reliability, and then move to validity if there are any other concerns related to the thread of validity.

So, Christie, I'll turn it to you to see if Terri wants to kick us off.

Co-Chair Teigland: Yes. I think as you pointed out, one of the biggest issues here that I think everyone who reviewed this measure on Subgroup 1 pointed out was the concern about under reporting of the bloodstream infection.

So, let me turn this over to Terri to discuss that relative to the reliability vote. Terri.

Member Warholak: Hello, everybody. And it's almost noon here on the West Coast. So, it's still good morning. So, I'm going to start us off on the reliability discussion.

But I have to admit, this is kind of a difficult line to discuss reliability on. Because reliability testing really wasn't done. So, I'm not really sure what to say.

And however, I think there's a lot of questions that the other reviewers brought up that I'd like to pose to the group, to get some thoughts and feelings on that.

And so, first of all there was a suggestion by some of the reviewers that there definitely is a missingness. It's something that we need to talk about, specifically for reliability. How can we rank, and order, and put groups into categories if we can't be sure that we have the data?

Also too, if there's under reporting, like a question I had was, if a facility does have under reporting, do they know that they have under reporting? Or does their measure actually look okay? I don't know. So, that's something else to think about.

There was also questions about how the degree of under reporting might change the rating of the facilities, as mentioned. There was also questions about reporting of the kappa agreement, and the consistency of it said, reviewers said, the data element reliability, validity didn't consistently report kappa agreement.

And that specifically that vascular access type, for example, used overall agreement. So, that's another thing to think about. And some others asked for a discussion of what use will the measure be put to? And is there a comparative or ranking component to use? Again, the impact of the under reporting.

And, let's see, and I think those were the major issues. But I'll stop talking at this point and see what the group has to tell us about that.

Co-Chair Teigland: So, Subgroup 1, any responses to Terri's questions? All good questions, Terri, about this measure. Coming from manual records. Ron, if you're talking, you're on mute.

Dr. Pickering: I also see Paul with his hand raised.

Co-Chair Teigland: Paul. I can't see hands. So --

Member Kunisch: Yes. This is Paul Kunish. Just to reinforce some of the questions that are raised. Whenever you have missing data you can't make a presumption the data's missing at random. So therefore, the data that you collect may not be representative of the entire picture.

And here you have the added problem of under reporting may actually be reflective of sites that are poor performers in one way or another. And therefore, they would actually be rewarded for being poor performers by under reporting.

So, it really gets, when you have this degree of missingness of data it really gets into some serious issues.

Co-Chair Teigland: Yes. I think several reviewers have pointed out the non reporting at the facility level. And that seems to undermine the validity. And again, reliability was kind of based on the validity testing here. Anyone else have comments on reliability?

Ms. Benin: Christie, this is Andrea Benin from CDC. Just let me know when you want to pass the --

Co-Chair Teigland: Yes. I think, I don't --

(Simultaneous speaking)

Co-Chair Teigland: Yes. I don't see any other comments right now. So, let's go ahead with, yes, the developer. Please go ahead with your responses.

Ms. Benin: Thank you. Thank you. And thanks for having us, and for, you know, really the hard work and dedication, and the digging into the nuances, and the trials and tribulations of this measure, obviously, as well as other measures.

And I think, you know, the context that Paul and Terri have described is important here. I would like to try to lend a little bit of broader context to this.

And I will, let me just see if I can figure out, if I can just share super quick. I've pasted a couple of the items. I don't know if you can see this PowerPoint.

Co-Chair Teigland: Yes.

Ms. Benin: But I did --

Co-Chair Teigland: We can.

Ms. Benin: -- just paste in, this is just some of the tables from the materials that I think you have that I just pasted here. Just to highlight a particular aspect of this that I hope you can take into consideration.

I mean, I will say that, you know, one of the things that has slowed down our validation obviously in recent years is the amount of disproportionate time has been spent on, you know, COVID-19.

And so, in the past few years in preparation for this maintenance we would normally have done, you know, some more, you know, additional work around validation that has just honestly not been possible.

So, you know, we recognize that some of the information that you're presented with is a little bit limited. But I would like to highlight a particular lens to this.

So, these are tables that are in the materials you have, except for the fact that I have added this column here to these tables that is called approach, just to highlight that really the bulk of the information that we are working with here for looking at reliability is what we call targeted validation.

And that's targeted because it is designed to optimize the use of resources, right. These are really big projects. It's very hard to have, you know -- and the work that CMS does to do validation reliability in an ongoing fashion is to target what is believed to be under reporters.

So really, to highlight what Paul said, which is that the charts that are reviewed for this purpose, and the facilities that are reviewed for this purpose are the ones that appear statistically to be likely to be under reporters, okay.

So the, and there is a little bit more detail in the methods that I think are presented in the packet. I don't remember if we've included all of those methods in there.

But the gist of the sampling is that it involved multiple strata. One of the strata is those who reported no cases, right. So, you didn't report any cases. So, you might, you have higher likelihood to get sampled.

And then there's another stratum that has facilities lower than their predicted, right. So, the measure afoot here is an observed to predicted ratio. And if you're observed as much lower than you're predicted you're also likely to be sampled.

So, the idea here, and this is why it's a little bit tricky around how you want to think about this in the context of reliability. But that for resource utilization the most important thing is to really find places that might be under reporting.

And that's important both from kind of a regulatory perspective and CMS's lens. It's also important from our perspective. Because if we're thinking about performance improvement, again, as Paul said it's the folks who, you know, who are likely to be, you know, maybe wanting to perform better without actually being there, who might purposes under report.

Or it could be the ones that are accidentally under reporting, you know. I don't want to cast aspersion on anything here.

But so, from the kind of scientific side of this, most of the validation that we have here is intended to really enrich for under reporters. And so, it did that. And I think that makes it very hard, you know.

And it's a little hard I think sometimes for us to explain that. But I'm hoping that some of the materials here can, you know, help you as you're evaluating, you know, the approach.

And since this is a maintenance activity for us, it's very important in a maintenance activity that we're optimizing the resource approach. So, I'll just say that as far as the bulk of this.

And similarly for the state level reliability, those are also typically targeted approaches to the reliability testing. So, you know, I'm hoping that that understanding can clarify. I can come back to this other part when we get to validation.

But just to give you a little bit of flavor for why that data looks so particularly, you know, odd. And those facilities, to answer Terri's question, you know, there is a lot of follow-up that happens here.

And within the NHSN application there, the facilities are able to see their observed rates, their predicted. They're able to see the difference between those.

There's all kinds of what we call TAP reports, which are targeted reports for identifying opportunities for improvement essentially. There's a lot of infrastructure around helping facilities to see how to handle potential improvements in this space.

So it, I'm hoping that facilities are not blind to this. It's a lot of effort to going into having them not be blind. So, they should know. And certainly if they have problems passing their CMS validation I suspect that they know that. And hopefully that's one.

Were there other aspects of things that I could clarify here? Sorry if I put this black box. I'm looking to see if anyone -- Those seem to be the key things that needed clarification.

There's probably a couple of other things in the report that I could try to speak to if that's helpful. But I don't want to keep you guys, I want the West Coast to get their lunch. So, I don't want to keep you too long.

Co-Chair Teigland: Terri's happy. She's got a little break there to get a couple of bites of food.

So basically you're saying, you know, you're really targeting for this validation those facilities that you expect to have underreporting. So this is sort of an overstatement of the over -- is that fair to say?

Ms. Benin: Right. It's an enrichment. I mean, part of the problem is we can't necessarily state. And I think one of the things that we can do is try to reach out to the group that performs some of this and get a little bit more information on how the strata break down.

And we can do that for our full submission if that's helpful. If you guys are going to -- you know, if you think that would be helpful, we can add that information. But the numbers will also be very small. So it's going to be very hard to create generalizable information there.

And I think Jonathan Edwards, our statistician is one the phone, and he may have some other ideas about that. We may be able to get a little bit of an understanding of how that breaks down between those that have zero events and those that have fewer events than predicted.

Was someone raising a hand there? It looks like maybe there was a hand up.

Dr. Pickering: Yes. I see Patrick's hand is raised and Alex. I think we'll start with Alex first since he's in a subgroup one. And then Patrick, we will come back to you when we open it up for the other subgroup members.

Member Sox-Harris: Thank you. I just wanted to comment that, you know, I share this concern about the underreporting. And I appreciate the clarity that some of the data we are looking at was targeted to those who were expected to have higher rates of underreporting.

But even looking at the earlier year random selection, although those are a smaller sample size, I'm less concerned about the overall percent of underreporting. I'm concerned about the facility level variation because that really gets at your ability to compare apples to apples.

So, for example, if the overall underreporting rate is 32 percent and everyone is at 32 percent that means something different than if many are at 5 percent and some are at 75 percent. So without understanding the variation in underreporting, it's hard to know what we're looking at. Thank you.

Co-Chair Teigland: Do we see any other hands on subgroup one? Matt, I don't.

Member Warholak: Patrick has his hand up, oh, but he's --

Co-Chair Teigland: Yes. He co-chaired for it, right?

Dr. Pickering: I don't see anyone else unless anyone has anything to share.

Co-Chair Teigland: Let's go to Patrick. We'll let subgroup two have the floor back.

Member Romano: Okay. Well, I'm not voting here, but I just want to perhaps ask a question for NQF staff to clarify or the committee chairs.

So this is a maintenance of endorsement situation, which means that we're looking for testing at both the patient and the counter level and the accountable entity level.

And if I'm reading the form correctly, there is no accountable entity level reliability testing in terms of understanding how much of the variation across facilities is random variation versus variation attributable to a quality signal at the facility level.

So could you clarify, does evidence of validity as presented here qualify for accountable entity level reliability?

Dr. Pickering: So not for -- so for the data elements level testing, that would apply to reliability but not accountable entity level. So at the data element level, the validity testing is done. That is applied to reliability testing at the data element level. It wouldn't apply for accountability entity level. Does that answer your question, Patrick?

Dr. Romano: I think so. So it suggests a deficiency in the presentation here unless I'm confused, but I defer to the Subgroup 1.

Co-Chair Teigland: No, I think that's right that the accountable entity level validity was not really documented. Except they did point to declining values over time of improvement, probably due to the extensive programs you were describing. Any other thoughts on that?

Dr. Pickering: Yes. I'll add here. For maintenance measures, we would like to see standard testing, which is usually at the accountable entity level. You know, we would like to see that. But for the most part if they have done previous testing with data element level testing, that is still fine except you wouldn't be able to get a high rating. You know, the max you could get is a moderate rating for just data element level. So just wanting to mention that.

Yes, for maintenance measures we would want to see expanded testing, whether it would be expanding the population or whether it would be at a different level, like the accountable entity level, but it's not necessarily required. As long as they're doing the empirical testing, which they have presented, they can also present prior evidence to support that the testing still is sufficient.

Co-Chair Teigland: Measure developer, any other comments?

Ms. Benin: No. I can just say that the data element validation that we did this past fall and, I guess, summer in preparation for the measure maintenance activity was around the critical data element of vascular access type.

So the way that this measure is constructed is that the -- you know, again, it's an observed to predicted ratio. And the predicted is calculated by multiplying essentially, you know, by taking each of the vascular access types and developing, you know, based on the patient months of each of those and the corresponding national rates during the baseline year, it creates a predicted number by vascular access type and then those are added together to create the overall predicted.

And so that data element was validated, and we were able to demonstrate high concordance between the manual, you know, the kind of re-abstraction manual chart review and -- did we lose everybody?

Dr. Pickering: No. We're still here.

Ms. Benin: Okay. Sorry. All of a sudden everybody went blank and something said we're having trouble connecting to the network in the upper corner.

And so between the manual chart review and, you know, what we have as reported to us. And so that effort was to make sure that that data element was functioning, you know, in the way that we would expect.

So as far as data element validation, you know, that particular data element is important to the construct of the measure. And in addition to that we took as a parallel activity a statistical modeling exercise whereby we re-evaluated if it would be appropriate to think about this measure in constructing the predicted count using a statistical model and using other factors in that model beyond vascular access type.

And the result which you have in the packet there really indicate the vascular access type remains, you know, by and large the important factor to be included in constructing this measure.

And so for that reason, you know, we decided not to move forward right now with any changes to the construct of the measure. So the measure remains constructed in the way that it has been. And so those were the two things that we did in order to reassess, you know, that critical data element.

Co-Chair Teigland: Okay. Thank you.

Ms. Benin: If that's helpful.

Co-Chair Teigland: And we'll discuss, I think, a little bit more about the risk adjustment when we talk about validity. But, Larry, I see you have your hand raised and then we'll go to Sam. You're on mute, Larry. Trying to get off mute.

Member Glance: Thanks. Hi. Thanks. So I understand what the measure developer is saying. But I think in terms of data element validation, I think it's critically important to validate not just the key variable that's being used in the risk adjustment but also the outcome element.

And I also understand what they're talking about in terms of targeted validation. But unfortunately for this particular measure, data validation is the key thing that the developers are using for both reliability and validity testing.

And because of the way that validation has been done, namely targeted validation, the results suggest that the data element is not valid. And that's really a big problem, I think, in terms of going forward with this measure.

And I'll save my comments on validation until we get to that point. Thank you.

Co-Chair Teigland: Yes. Thank you, Larry. Those are good points. Sam.

Member Simon: Yes, just, I guess to piggyback a little bit on what Larry was saying. One other thing to note, I guess, about the data validation results here is that from my understanding of the package, there was no chance adjusted agreement. It was purely raw agreement. And my understanding, too, is that NQF requires kappa or chance adjusted agreement. So I think that's something to note as well.

Co-Chair Teigland: Yes. Good point. Anyone else? This is a bit tricky since we're voting reliability based on the data element validity testing. But if there's no other comments, maybe we can vote on reliability and then move to the validity discussion, which I think has some broader issues. Does that make sense? Gabby?

Ms. Kyle-Lion: All right, everyone. I will go ahead and share my screen and open up voting. Just as a reminder, this is only for Subgroup 1, and there are no recusals for this measure.

Okay. Voting is now open for Measure 1460 on reliability. Your options are A for moderate, B for low and C for insufficient. I believe we should have nine people for this vote. So we're looking for nine votes. I'm seeing eight so we're just waiting on one more.

All right. We're at nine. We'll go ahead and close the voting. The voting is now closed on Measure 1460 for reliability.

There was one vote for moderate, three votes for low and five votes for insufficient. Therefore, the measure does not pass on reliability. I'll pass it back to Matt and Christie.

Dr. Pickering: Thanks, Gabby. I'll turn it back to Christie.

Co-Chair Teigland: And I'm going to pass it right off to Jeff to really talk about the validity testing concerns.

Member Geppert: Yes. Actually I'll go kind of quickly because a lot of the things that I was going to mention have already been discussed.

And to my mind, this was really kind of an application of the NQF criteria if you look at -- I think it's really about sort of the sufficiency of what was presented.

So for data element validity, as Larry mentioned, it's supposed to be for all the critical data elements. And as Sam mentioned, the flowchart, you know, says that if you don't provide sort of a kappa statistic than the proper weighting is insufficient. And the guidance that NQF provided us before the meeting also says that, you know, simple percentages are not acceptable for data element validity.

So from that perspective, it just seems like insufficient is the right sort of rating. But I think there are other aspects of insufficiency that we've already kind of touched on.

The data validity analysis that was done has been described by CDC. You know, it was done on this targeted sample. The intent was to identify sort of low reporters based on low reported rates. That's really sort of more of a program monitoring kind of activity not so much intended to sort of assess data element validity.

But at the end of the day, and the developers sort of responded this way in their comments, that we're not able to make any kind of inference based on the data that was presented to the larger measure, that no inference is really possible or even necessarily appropriate.

But that's what we're doing here is we're making an inference about data element validity for the larger measure. So we can't really make the appropriate inference based on the data that was presented. So that's sort of another respect in which the information presented is not sufficient. Although as the developer mentioned, there is a much more robust analysis of the vascular access.

And I think the third way in which the information presented is insufficient is sort of in the interpretation. Even in the results that were presented, there is sort of a lot of speculation about what the potential causes are of the large number of missing reported data.

But a lot of the sort of hypothesized causes are really more reliability sort of issues. It's a better understanding of the specification, more training related to the specification, you know, changes of vendors that could potentially have resulted, you know, inconsistent application of the protocols.

Those to me are more sort of reliability issues because they speak to better training, better protocols, better data collection unless, you know, the entities that tend to be underreporting, you know, or there's some sort of systematic relationship between that and the lack of understanding and the lack of application of sort of protocols but that's not really sort of investigated or described.

So I think there are several respects in which what was presented is kind of -- we're really not sort of even able to make an assessment of the validity of the measure based on what we've been given.

Co-Chair Teigland: I'll turn it to Larry. You have your hand raised. You're probably going to talk about the risk adjustment, I suspect.

Member Glance: Sure thing. Hi. So I wanted to piggyback on Jeff's comments. In terms of one of the things that I think we should be also thinking about when we're evaluating validity is the threat to validity, which is the risk adjustment model.

And I would suggest that although we have agreed on the chat and in various discussions not to spend too much time talking about individual pieces of the risk adjustment model, I think in this case, there is virtually no risk adjustment model. And I think bears mentioning.

The only risk factor in the risk adjusted model is the type of vascular access. And to me that really lacks face validity. The risk of infection in these patients is multifactorial. And as a clinician, it involves whether it's a medical or a surgical patient, what type of surgical procedure, a CABG, for example, versus a colorectal patient. For medical patients, is it somebody who was admitted with congestive heart failure or was it somebody who was admitted with sepsis, I mean, just to sort of name a few. So really, you know, I think to call this a risk adjusted model, it's very insufficient.

The other thing in terms of the testing of the model, what they did is they tested their model against the null one basically. Risk adjustment versus no risk adjustment. They didn't look at any other aspect of model performance. So I think that is also -- that's not adequate.

So in terms of threats to validity, I think that risk adjustment -- the lack of appropriate risk adjustment and the lack of appropriate testing, I mean, is very important to consider.

Co-Chair Teigland: Yes. Thank you. Let's hear from Paul.

Member Kurlansky: You know, Larry raised it. I'm glad he raised the question as we knew he would. But this is sort of a question for our method panel chair people, and it relates to the discussion in the chat.

But if you have a situation where as a clinician the risk model or the absence of the risk model, as the case may be, itself does not achieve face validity, in other words, it just seems so clinically inappropriate not to more carefully risk adjust in this particular case, is that within the domain of the scientific methods panel or is that clearly off limits and something for the Standing Committee only?

Co-Chair Teigland: I think it's within our domain as long as, you know, we are looking at the scientific acceptability of the risk adjustment approach. That's part of what we evaluate. Matt, am I saying that correctly?

Dr. Pickering: Right. So taking into consideration the approach that was taken, if the approach seems to be rigorous and then just, you know, adjusting, you know, the other aspects of the model, like the C statistic, discrimination, things like that, right? Things around the type of factors to be included, we want to try to leave that for the Standing Committee's input and decision-making that includes the different types of social factors. But if the approach determines inclusion or non-exclusion looks sound, that's what the SMP is being asked to evaluate.

Co-Chair Teigland: Larry, a comment on that?

Member Glance: Yes, I was just going to -- I think this is a really, really important discussion. And we've been having this through the chat, but I wonder if we should be having it more as a committee. I think that there -- as Paul mentioned, there are cases where the risk adjustment is just so problematic that even if the committee, the methods panel, doesn't always have the full content expertise that the Standing Committees have, I think it's still important for us to be able to look at the validity of the risk adjustment model as a whole, not just in terms of the mechanistic pieces, like the C statistic and calibration and things of that type, but also looking overall at the model how some of those factors are specified, whether they're mis-specified. We ought to be looking at those things.

I think there are cases, for example, where a Standing Committee might not understand that certain risk factors should possibly be specified maybe as a categorical variable because association between the outcome and the risk factor is non-linear as opposed to linear. I mean, there are all sorts of things that I think that our committee, our panel should be looking at.

And the other piece that we have discussed in years past, but we haven't talked about recently, is that if in fact the risk adjustment model is problematic or if there is a lack of risk adjustment, that bears on the reliability discussion.

And I think, David, we've had this discussion, you and I, many times. But if you have poor or inadequate risk adjustment then you're going to see much greater between differences for the facilities so that your reliability will be artifactually elevated.

So, again, looking at the risk adjustment model not only impacts on our ability to look at the validity of the measure but also to look at the reliability of the measure.

Dr. Pickering: Yes, thanks, Larry. I'll just add that it's not to say, you know, that SMP shouldn't take into consideration, you know, the fact these are included and the different type of parameters around those. You know, if something to where, you know, the Standing Committee has some input on related to is this factor clinically relevant? Can this factor be meaningfully influenced by the accountable entity?

Those types of discussions we reserve for the Standing Committee. And the concerns around how the developer approached those factors, inclusion or exclusion, will be documented from the SMP and shared with the Standing Committee for decision-making.

I know there's still a gray zone there. But if we can try to keep that in mind as -- you know, moving these decisions around, factor inclusion or exclusion, if the approach looks sound to more of a Standing Committee consideration, that's sort of where we're trying to hang our hat on. I appreciate the comment from Larry.

Co-Chair Teigland: Well, Matt, are (simultaneous speaking).

Ms. Benin: I'd like to --

Co-Chair Teigland: Go ahead.

Ms. Benin: Oh, sorry, Christie, go ahead.

Co-Chair Teigland: No. I just wanted to say that our review process does ask us to evaluate whether the risk adjustment model was appropriate. Did they, you know, look at the rationale for including and provide explanations for not including risk factors? It is presented to us that way as part of our evaluation of, you know, the appropriateness of the risk adjustment. But, yes, go ahead, CDC.

Ms. Benin: Yes. I'd like to maybe clear up some misconceptions that were just presented. I think for starters this is an outpatient dialysis measure, right? So this is a person who is receiving dialysis at an outpatient dialysis facility.

And I think just to make sure that we have that context in place as far as what may or may not present itself around base validities or not post-surgical, immediate post-surgical CABG patients or some such thing, these are folks who are receiving outpatient dialysis for their end stage renal disease.

And in addition, the approach that we have taken for risk adjustment is identical to the approach that we have taken for all of the health care associated infections. You know, this is in parallel with, you know, a handful of other really quite well-accepted measures. And so the concept of how we think about, you know, face validity in that construct, I think, should be put in the context of how the field is measuring and adjusting for health care associated infections occurs.

What we did specifically for the purposes of this maintenance application was to rejustify, essentially, the use of what is not truly a risk adjusted approach to calculating the predicted, but what is in fact, I think, I will call it more like a risk stratified so we stratify and then we add it back up. So it's not technically a risk adjusted model, but it is in effect the same mathematically.

And so we relook using a statistical modeling exercise. We relooked at all of the possible variables that we can put into that model and reevaluated if any of them should be in there.

So we completely attempted to reconstruct the model and found that it was not superior to the existing methodology. The existing methodology is far simpler and enables facilities to perform their own calculations.

And so we decided to stick with the existing methodology. This is a measure that's been in play for, I don't know, five to seven years. So we decided to leave it with the existing methodology.

If we were to decide to want to move into additional patient level risk adjustment, which we could all argue if we step back and really think about what we're trying to do here, which is prevent these infections so we could argue whether -- and I think we could probably argue for days and write several thesis about it, as to what really should you adjust away in these settings, right?

So these are, you know, infections. And so we could conceivably think about in the future measuring patient level risk. That, however, does require the submission of all of that data for the denominators as well, right? You cannot just collect that on the event of termination.

And so it's a much bigger and somewhat unseasonable and impractical list at this time. And we are working towards that for sure. But in the context of this measure and the necessary face validity for this measure, it's clearly not feasible right now.

There may be additional things in the future. And, you know, we may be back here another day, you know, with some magical data from these LVOs. But that's a different story. And so for the purposes of this and what is, you know, what presents itself as face validity the other factors do not stay in the model. And Jonathan can try to explain that better for folks if necessary. But this not for a lack of looking at those adjusters. So I just want to make that part clear.

Co-Chair Teigland: Yes. Thank you for that clarification.

Ms. Benin: If you need more description at the statistical level of the approach of that modeling and the bootstrapping and calibration statistics, et cetera, et cetera, you know, we can have that. I don't know that everybody wants to sit through that now. And if we didn't reflect that adequately in the documentation, you know, we can certainly work on that.

I don't think we realized that having the kappa statistics was an essential NQF check box. We can calculate the kappa statistic to add that to the materials. So thanks for highlighting that as well.

Co-Chair Teigland: Yes. Thank you. I think we are required to vote on the information we have at hand when we're ready to vote. So that's where we are right now. But certainly for the future, I'm sure, you know, if you can submit that information. It might be helpful.

Last call for any other comments on this one before we move to a vote on validity. Seeing none, I'll move to Gabby.

Ms. Kyle-Lion: All right, everyone. As a reminder again, this is for Subgroup 1. And we have no recusals on this measure. Okay. The voting is now open for Measure 1460 on validity.

Again, I believe we are looking for nine votes here. We're at eight. Just waiting on one more. All right. We're at nine. The voting is now closed for Measure 1460 on validity.

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

Co-Chair Teigland: You said we're going to take a break, Matt?

Dr. Pcikering: Yes, I did. It's probably only going to be about five minutes if that's okay just because we still have two measures to go through. So I want to thank our CDC colleagues as well for their time in presenting the measure and answering questions from the SMP members.

But if we could just take five minutes. We'll come back at 3:45 on the Eastern side, so 3:45 p.m. just so we can get back to the two measures. And we'll start off with 0471E. A five minute break.

(Whereupon, the above-entitled matter went off the record at 3:39 p.m. and resumed at 3:45 p.m.)

Dr. Pickering: Okay. I have 3:45 p.m. on the Eastern side. I appreciate everyone coming back from the short break. We're going to see where we get today. We have two more measures. If we are not able to get to the last measure, 0716e, we may have to move that to tomorrow. But we'll try to see if we can get through both of them.

So both of these measures, the developer is the Joint Commission. So I just want to check in, do we have members from the Joint Commission on the call?

Ms. Walas: Good afternoon. This is Chris Walas from the Joint Commission, and our team is ready. Thank you.

Dr. Pickering: Thank you so much, Chris. Okay. And then Christie, are you back on?

Co-Chair Teigland: Yes, I'm here.

## 0471e ePC-02 Cesarean Birth

Dr. Pickering: All right. Great. So, Gabby, I'll go ahead and ask you to pull the slide up. We'll start with 0471e. You can see who are lead discussants there are, Sam Simon, lead and the secondary is Paul.

And for this one, I'm going to rest my vocal chords and give to Hannah to introduce this measure. So, Hannah?

Ms. Ingber: Thanks, Matt. Yes, so I'll just go briefly over the testing that was presented and some information about the measure.

So this is a new outcome e measure that assesses the number of nulliparous women with a term singleton baby in a vertex position delivered by cesarean birth.

This measure uses electronic health data to measure at the facility level. It is not risk adjusted.

This measure did not pass preliminary review on reliability or validity. Testing presenting for the encounter level validity testing serves as the reliability testing. And the developer's data set included two pilot sites for a total of seven hospitals using either Epic or Meditech EHRs.

So jumping to the encounter level validity testing that was presented, staff manually reabstracted each data element from the EHR and blinded and compared that data for agreement and accuracy.

So the developer used several tests to analyze that agreement accuracy. The measure outcome rates and data elements were tested for specificity, sensitivity, positive predictive value, negative predictive value and agreement rates with kappa.

Specificity and NPV were considered high for Sites 1 and 2 ranging from 82 percent to 100 percent. Sensitivity and PPV were considered high for Site 1 but low for Site 2 where PPV and sensitivity were both 0 percent.

A final plot was used to test meaningful differences and the ability to detect outliers and significant variation in measure rates.

There were differences in measure rates across the pilot sites ranging from 78.9 percent to 96.5 percent. The developer reported that without exclusions, measure rates increased overall by 17 percent or 4.7 percentage points. Exclusion rates ranged from 0 to 16 percent.

The developer did not risk adjust the measure and offered the following reason. Exclusion criteria were chosen to ensure that the target population would be women with nulliparous term singleton vertex or NTSV pregnancies who have a lower risk of maternal morbidity and mortality during a vaginal birth delivery than do women who have undergone a previous C-section. Therefore, the population of women in the denominator as a result of the exclusions allowed the measure to focus on a more homogenous group of women where the greatest improvement opportunity exists as evidenced by the variation in rates of NTSV cesarean births indicating clinical patterns may affect this rate.

So as a reminder, when a measure developer uses encounter level validity testing to demonstrate reliability as we've seen many times today, your vote on reliability should rely on that encounter level validity testing. I'll pass it to Christie now.

Co-Chair Teigland: All right. So there were quite a few concerns expressed by the SMP members that reviewed this measure about reliability. I'm going to pass this to Sam to start that discussion off and lead that discussion. Sam?

Member Simon: Sure. I'm happy to sort of kick this off. I'm going to focus a majority of my comments really around the data element, validity results that were provided by the developer and focus on sort of, you know, what I think is somewhat of a technical issue with the results as well as sort of a larger feasibility issue I think that threatens the validity of the measure.

I'm going to touch on the risk adjustment issue. But I think the punchline is, I think, I'm sure people will weigh on this, but I do think we want to pass that along, that issue to the Standing Committee.

So in the original submission, the developer did not provide kappa agreement results for the two denominator exclusions, which are placenta previa and abnormal presentation. And given that some sites actually had 15 percent of the denominator cases excluded because of one of these conditions, this was a pretty important omission. And, of course, NQF guidance dictates that if data element validity is evaluated, all key data elements should be included and so this is needed. And so there was an agreement, but it wasn't a kappa agreement.

So the developers did respond to this omission. They did provide a table of kappa agreement rates for all of the measures' data elements. And that table is Pages 39 to 40 of the discussion guide that we got.

But what's concerning, and this may be more of a technical issue, but what we note is that the overall agreement rate, or what's listed in the table as a match rate, for every data element is exactly the same as the chance adjusted kappa agreement rates, which to me strikes me as not plausible. And it does actually call the results into question. I think this is less of an issue for Measure 0716, the complications measure.

But ultimately, those results in the table that the developer provided on Pages 39 to 40 of the discussion guide, to me really raises more questions than they answer. And I would really recommend that the developer in the future show all values from the cells of those 2x2 tables that are used to generate kappa agreement if you're relying on data element validity results.

But the other thing I'd like to just point out is that so one of the test sites, Site Number 2, clearly had some real difficulty reporting the numerator for this measure.

The hospital used Meditech, which is a pretty commonly used EHR system. I think it's like the third largest in terms of their share of the market, which means that this could be an issue for a substantial number of hospitals who could report this measure.

So this reliability issue -- I'm sorry this feasibility issue on reporting gets to validity and reliability. It sort of underscores sort of the things that we get concerned about in terms of reliability and validity because if the data aren't collected in structured fields as the developer described, it's going to be really hard to evaluate the measure.

You know, as the developer stated in the original submission on Page 26, this hospital uses a standalone OB documentation system that doesn't interface with the EMR and that the OB documentation in Meditech is in non-discrete fields in a PDF form.

And I'm completely familiar with this issue but for an

ECQM, this presents real problems. You know, I do understand that -- you know, the developers do explain that changes have been made at the site, at this particular test site. But, again, we sort of have to go with what we've got.

Just a couple of other points to make, some things that gave our subgroup pause is just that the data results are -- the results are based on a total of 123 births across seven hospitals. So we're not talking about a particularly robust sample. And then, of course, the fact that this is an outcome measure without risk adjustment, it's just always going to raise eyebrows on this committee.

The developer did provide a rationale for the lack of an approach. And, again, I'm going to defer to the clinicians in the Standing Committee to take this on.

But just given the aforementioned lack of clarity around kappa for key data elements, including exclusions, the substantial difficulty for a large EHR system to collect the data for this ECQM and the small number of cases in the sample, the data provided didn't give me a lot of confidence that this measure is reliable for valid.

I do want to say that, you know, what the developer did is completely acceptable in terms of the methods they followed, in terms of looking at a comparison between EHR extracted on the chart abstracted data to assess reliability and validity. But I do hope that in the future we as a panel can sort of reconsider whether data element validity can or should obviate score level testing for reliability.

I can really imagine a scenario where there's great agreement between an EHR and manual abstracted data. But provider variability results in a measure score with pretty lousy signal-to-noise reliability.

So let me stop there. Paul, I'll turn it over to you to see if you've got anything to add here.

Member Kurlansky: Yes, I think you hit all the major points. Just to, you know, emphasize if the methodology for establishing data element validity is correct, but the results are extremely disturbing, then that goes to the core of the potential reliability of the measure.

And the fact that, you know, one -- it was only tested in one hospital so I can't say for sure, but I think we should find out whether all Meditech or just this one particular hospital's Meditech system didn't do this, but the fact that there was a mitigation strategy is very nice. But you can't expect there's going to be a mitigation strategy for every single hospital in the country that doesn't have the EHR capability to participate so, you know.

And certainly Meditech is, you know, 17 percent or so of the EHRs in the country. So even best case scenario if 17 percent of hospitals can't participate in an effective way is deeply concerning, and I think I will leave it at that because I think he really brought out all of the other major issues.

Co-Chair Teigland: Mm-hmm. And there was no testing of the within hospital variation as well. Larry, you have some points?

Member Glance: Yes. So, yes, I think that I just want to echo the points that were made earlier. I think that the lack of reproducibility of the outcome variable, whether or not a patient actually underwent cesarean delivery, I think that's critically important. You know, I think none of us would pass a mortality measure if the mortality outcome was not accurate.

The second piece is the lack of risk adjustment. I understand the measure developer's point that if you make this a supposedly homogenous low risk population that you will be able to potentially not need risk adjustment. I'm not sure that the population was homogeneous enough to really justify the lack of risk adjustment.

And in particular, I think that there are certain risk factors that would be associated with a high risk of cesarean delivery that would not be excluded by the exclusion criteria that we used in this measure for things like advanced maternal age, high BMI, certain obstetrical conditions for long labor, things of that type were not included and were not excluded. I hope I'm right on this one. I apologize for the mistake I made on the prior one.

They did say that the rates of cesarean delivery were similar in university hospitals compared to what would be expected of a sicker case mix than non-university hospitals. And they used that as a justification for not using risk adjustment. I'm not sure that that's enough of a justification.

And finally I would like to bring up a point that we typically don't consider when we're looking at outcome metrics, like mortality or complications that may be appropriate in a case like cesarean delivery. So in theory, in practice, we always want the complication rate to be as low as possible. We always want the mortality rate to be as low as possible.

And, again, this is may be something more for the Standing Committee, but we don't always want the cesarean delivery rate to be zero. Okay? We don't want it to be too high, but we also probably don't want it to be too low because if you never do a cesarean delivery, there are cases that are appropriate for cesarean delivery that are not being done.

Again, that may not be something that we, as a committee, want to grapple with. On the other hand, it is something that -- this is not a simple outcome metric because the lowest values may not necessarily be the best value. So the typical approach, which is to use folks who use facilities that have the lowest possible outcome rate as the highest quality facilities may not be appropriate here. And that kind of goes, I think, as a threat to validity.

Co-Chair Teigland: Yes. And it seems like the threat you just described might be mitigated if there were appropriate risk adjustment to those rights but, again, that may not be within our purview.

Any other comments from either Subcommittee 1 or Subcommittee 2? We can open it too, before we hand this over to our developer for comments. I don't see any hand. Do you Matt? Chris, do you want to respond?

Ms. Walas: Hi, yes. This is Chris Walas from the Joint Commission. Thank you. We appreciate the feedback. And I'd like to speak to parts of it and then I'll call on my colleague, our statistician, Stephen Schmaltz, to discuss the kappas when I finish.

So first of all we do recognize the concern for validity of this measure. However, as the measure developers, we would argue that the measure is valid when other supporting data is evaluated.

As you know, only two sites volunteered to participate in pilot testing. And we realized early in testing that the issues at Site 2 could impact our overall reliability and validity results.

As a matter of integrity and transparency, the results were included in our study. And we countered these results at this one small site with the following. So the overall lower kappa levels for Site 2 were actually due to 10 specific data elements.

Issues with three of those data elements we feel are resolved as the current version of ePC-02 has updated author date time to relevant date time and allows two ways to determine gestational age calculated using date of delivery and estimated due date or reported EGA.

These data elements are shared with ePC-07, and ePC-07 pilot testing shows excellent metrics in three EHRs, Epic, Cerner and Meditech with 94 to 98 percent match rate.

The other seven data elements are related to preterm and term parity results and their associated author, date, time and then gravida author date, time only. The gravida result actually had a 91.2 percent match rate at Site 2 and 100 percent match rate at Site 1.

So 57 percent of the mismatch for these specific data elements were due to missing data because of that issue we mentioned using the standalone OB documentation system that did not interface completely with the electronic health record, Meditech.

The OB standalone documentation was not Meditech. That was a separate, I believe, GE OB documentation system. We have, like I said, tested some of these data elements in ePC-07 and the Meditech system was able to accurately identify these data elements.

So the OB documentation that was present at Site 2 in non-discrete fields, as you mentioned, they did make a change to capture these data in discrete fields. However, they were unable to submit updated dated in time for our NQF submission. However, we feel confident that these data elements are able to be accurately extracted in an EHR system as evidenced by Site 1's 96 to 100 percent match rates out of all 10 of these data elements.

The ePC-02 measure logic only requires parity or gravity or pre-term and term, not all four of those data elements together.

And when accounting for the root cause, Site 2 had the low kappas which we, you know, were told they were able to mitigate. And using the other available evidence which shows high kappas and multiple EHRs for those shared data elements, we feel that overall this measure is valid and able to capture the differences in performance. And the chart based ePC-02 does correlate highly at .88 with the EC2M version.

I'll touch briefly on the risk adjustment. Since this is highly correlated with the chart base and, you know, we did a lot of investigation into some analysis, and there was an analysis of the SMFM proposed additions to NTSV exclusion code set that was taken in a manuscript in preparation. And I know I did present this in our documents.

However, we just want to point that adding the additional exclusions only resulted in that study in a .3 percentage point reduction. Another study that we looked at, that looked at age and BMI, these results indicated that physician preference and subjectivity accounted for most of the age and BMI affects, and NTSV cesarean rates would support the lack and need for risk adjustment.

And since we focus on this NTSV population, the measure is not meant to exclude all possible indications for cesarean birth. And we just want to decrease that primary cesarean birth in nulliparous patients because that's going to impact the overall C-section rate by decreasing the number of subsequent C-sections, you know, those planned C-sections because someone had a primary C-section and now they're going in for a repeat C-section or a C-section then that puts them at higher risk for future placental implantation issues which, you know, they may need a future cesarean.

So we're just looking to decrease the overall rate by focusing on this one portion of the population. And we agree that the cesarean birth rates, we don't know what the low level is. And we've made it clear in our specifications that we do not know what the lower level is and so while we used 30 percent as our recording cutoff, we do not publish the actual rates for any hospital on our site that is below that because we don't want to inappropriately suggest that someone should be lower because lower is not always better.

So we are careful about that, and we want to make sure that the variation in rates are -- you know, that there are standards and that these rates do lower as we go in the United States very high. However, we do not want to differentiate between those inappropriately low rates.

And I'd like to call on our statistician, Stephen Schmaltz, to discuss the kappa within hospital variation. Thank you.

Co-Chair Teigland: Thank you.

Mr. Schmaltz: Hello. Can you hear me?

Dr. Pickering: Yes.

Co-Chair Teigland: Yes.

Mr. Schmaltz: Okay. There were comments about not having -- initially not having kappas for the data element level, but it is questionable about whether kappa is really appropriate at that level. For instance, how did you define for a continuous measure what the chance abatement is? It seems like kappa would be the appropriate statistic in that case.

In fact for those kind of measures, it would seem like the match rate would seem to be the only one that could be used unless you used a correlation if you had two continuous variables. But how do you do it? You find it for two dates for instances. I don't think kappa would be an appropriate statistic in that case.

Now if you have people that report a measure or particularly data element or you don't report a particular data element, you can look and see how many of the original report and the re-abstracted date report, and you could calculate a kappa on something like that. But what about the data element that anybody collects? I just think you need more diameters from NQF on really what you should use other than match rate for the elements like that.

As for the variability, part of that variability is really due to the two obstacles that had trouble identifying the measured population, which was why they came up with a zero measure rate, which probably isn't the true measure rate. Any questions?

Co-Chair Teigland: No. Thank you. And I think the SMP reflected that in some of their comments that, you know, because the sensitivity was a function of the testing site, the differences in performance really couldn't be determined. It wasn't really differences in true performance.

But, Sam or Paul, do you have any responses to the developer's comments?

Mr. Simon: No. That's helpful. And, you know, I would agree. Kappa is not appropriate for a date of birth. I mean, that's clearly not what the intent is.

But I think the issue was more around sort of the total agreement and the kappa rates being exactly the same which, I've never seen that in practice. So that was a little concerning.

And I think it was really more the concern that there were some critical data elements that didn't have kappa rates in the original submission, particularly around the exclusion. So that's what that point was really about.

Member Kurlansky: And, you know, I think, you know, to the question of validity with the absence of risk adjustment, I think, Christine, you sort of hit the nail on the head is that, you know, rather than sort of setting an arbitrary, you know, 30 percent rate and then reporting underneath that, you might actually be able with more granularity get a good sense of what is appropriate by using risk adjustment.

Co-Chair Teigland: Are there any other comments? So, Matt, I guess, I'm --

Member Needleman: Wait. If I can, Christie. I want to give --

Co-Chair Teigland: Yes, sure. I didn't see hands.

Member Needleman: Sorry, sorry, yes, I didn't --

Co-Chair Teigland: Go ahead, Jack.

Member Needleman: I've been having trouble with the controls here all day including -- never mind. I want to give the developer another chance on this risk adjustment issue because the argument was we've used the exclusions in lieu of risk adjustment to basically deal with all -- get rid of most of the cases where a C-section might otherwise be expected or where there are other circumstances.

So, Chris, can you just walk us through quite explicitly the exclusions and why you think those are sufficient?

Ms. Walas: So the exclusions just get us to this NTSV population, which is what we're focusing on and why -- you know, the singleton is already in the measure logic with the codes. The vertex is the exclusion, so anyone who is not vertex would be excluded.

So the exclusions are just necessary to get that target population. And we feel that adding the other conditions that may impact a risk of cesarean birth is not what this measure is intended to do. The measure is just intended to look at this group of women to see what their C-section rates are.

And from the studies we do have a technical expert collaborator, Dr. Elliott Main, that has done a lot of research on this. And we worked very closely with him to make sure that there isn't a need for this. And this measure has been around for a long time. And we have had technical expert panels on this in the past.

And it was determined that adding more maternal exclusions just increases the burden and the complexity of the measure. It doesn't really impact the clinical intent. The clinical intent is to get these levels to a more appropriate rate.

And, you know, if you look at age and BMI, those, you know, may be distributed among hospitals. You know, they're not maldistributed in a way. So even if you risk adjust for them, you're still kind of at that same playing field.

And we know that. You know, the SMFM proposed additional maternal conditions to adjust for. And when they looked at that, it didn't impact the rates that much. It was a .3 percentage point reduction among all hospital types.

So to add that level of complexity and burden, we felt was not necessary when the clinical intent is to just get a decrease in this one population of women for their C-section because by decreasing that primary C-section you're increasing future repeat C-sections.

Now the other issue when you're looking at maternal conditions is sometimes, you know, when you're trying to add codes for risk adjustment, some of these codes are too inclusive of the condition where they would not necessarily need to have a cesarean section.

So to be able to include 100 percent of all of the conditions that could possibly result in a C-section would be very challenging, like I said. Some of the codes cover conditions that are very benign and some of them are very high risk so they or may not be the physician's judgment as to whether or not they wouldn't actually need the cesarean section.

So I'd like to call on Dr. Main. He has joined us, and he's done a lot of work in this. And I'm sure he can provide a better explanation to you all. Dr. Main?

Dr. Main: And I'll brief just because this doesn't seem to be the main issue for this measure at this time.

This is a measure that's widely used in the non-emeasured category and does -- as part of Healthy Person 2010, 2020, 2030 with a national target of 23.9 not 23.6.

We've done major quality improvement projects in California on this and have been able to reduce almost every hospital down into the mid to low 20s with this measure.

It is not really an outcome measure. It is the frequency of a procedure. So it makes it more of a process measure that has outcome implications. But it's not a measure of morbidity or mortality in itself. And there's certainly no reason to expect it to be pushed to zero or even a very low number.

In most usages somewhere in the low, mid-20s is a very reasonable number and that's what people are shooting for.

The attributable fractions for the risk factors is, as Chris said, these are negligible for anything else besides the ones that are in the exclusions with the exception of discussion of age and BMI.

Both of those are associated with increased C-section individually. But what's interesting is that hospitals that have high rates of high maternal age can have very, very different C-section rates in those populations. So it's not a fixed effect.

The same with high BMIs. There are some hospitals whose high BMI patients have very low C-section rates and others have high C-section rates. It's strongly implicating that it's the subjectivity of the provider that has a significant effect here together with the fact that hospitals that have high maternal age in their nulliparous populations also tend to have very low BMI and vice versa.

So hospitals that have younger populations tend to have higher BMI populations. So those factors in practice largely zeroed themselves out or canceled themselves out so that at the end of the day, you know, give or take a few percentage points, we're not dealing with much effects from those.

In particular if you look at studies that we've done if the same population were to deliver at a best practice hospital, which is just in the top half, top half for both baby outcomes and mother outcomes, you end up getting rates that are in the low 20s.

Member Needleman: Thank you. There's enough clinical complexity here, I'm inclined to let the Steering Committee experts deal with the clinical issues here. But if I can ask just one other question of Chris, which is I just want to make sure I heard this right, that the Meditech problem that we saw in your data, you believe that's not a systematic problem with the Meditech EHR or the Cerner EHR. That was just a one-off for the particular hospital that was in your testing group.

Ms. Walas: Correct. The OB system was not Meditech so it did not integrate with the Meditech system that they had. So it wasn't the Meditech that was the issue. It was the standalone system.

Member Needleman: Thanks, Chris.

Ms. Walas: You're welcome.

Co-Chair Teigland: All right. Last call for any comments before we move to a vote on reliability of this measure. Okay. Gabby, let's do it.

Ms. Kyle-Lion: All right, everyone. Give me a moment to share my screen. Again, a reminder, this is Subgroup 1. And I did want to note that Joe Kunisch is recused from this measure, but he is not in the subgroup so it should not impact voting.

With that being said, voting is now open for Measure 0471e on reliability. Your options are A for moderate, B for low and C for insufficient. And I believe we are looking for 10 votes here.

Dr. Pickering: Thanks, Gabby. And this is Matt. I just wanted to remind the folks who are voting that for reliability, you are using the data element validity testing to make your assessment of reliability vote so the results that have been discussed.

Ms. Kyle-Lion: We're still at nine votes. I'll just give it another minute. All right. We're still at nine votes, which is above quorum so I'll go ahead and close the poll. The voting is now closed on Measure 0471e for reliability.

There were four votes for moderate, three votes for low and two votes for insufficient, which means that the measure is consensus is not reached on reliability. I will pass it back to you, Christie and Matt.

Co-Chair Teigland: All right.

Dr. Pickering: Great.

Co-Chair Teigland: Yes. So I guess we move on to the discussion of validity and Sam and Paul you are taking these on.

Dr. Pickering: Oh, I hear a lot of typing there. Is there anything new, Sam or Paul, that hasn't been discussed? We've been talking about risk adjustment, et cetera.

But just because we're kind of getting a little bit close to the end of our time if there's anything new that you want to present or discuss, we can do that. If not in deference to the SMP members, we could potentially move to a vote on the validity test.

Member Simon: I have nothing further.

Member Kurlansky: And I have nothing further. We've spoken about it.

Co-Chair Teigland: Okay. Good. No new issues. No more comments. Let's vote on validity.

Ms. Kyle-Lion: Sorry. Give me one second to pull up my screen. Okay. Sorry about that. Voting is now open for Measure 0471e on validity.

Your options are A for moderate, B for low or C for insufficient. And again, we're looking for 10 votes here. And I apologize that the font is so small. I will make sure that doesn't happen again. All right. I'm seeing nine votes. I'll give another second to see if we get that tenth one. Okay. I'm still seeing nine. So I'll go ahead and close the poll because eight is quorum for this measure.

Voting is now closed for 0471e on validity. There were five votes for moderate, two votes for low and two votes for insufficient. And once again, the measure is CNR on validity. I will pass it back to you, Christie and Matt.

Co-Chair Teigland: Matt, I think we're going to push the next measure to tomorrow if I heard you right?

Dr. Pickering: That's right. So recognizing that we are close to the end of our meeting time today, we are going to move the measure to tomorrow. It will be the first measure we discuss.

So we're going to give it about 30 minutes for that discussion so keep that in mind. But for Subgroup 1 participants, we'll be voting on that or hearing the discussion there and be voting on that tomorrow. And so that will be the first measure up.

So with that, that does conclude our measure evaluations today. So I want to also thank the Joint Commission for their time today as well as our other measure developers that attended the meeting and discussed with our SMP members.

I'm going to ask the team to move to our public comments. So I'm just going to give some time to any members of the public if you'd like to comment on any of the measures that have been discussed today, now is your opportunity to do so.

We'd kindly ask that you use the raise hand feature on the Webex platform, and we can call on your accordingly. So, again public comment is open, and we'll just give it a couple moments here.

Again, this is an opportunity for public comment on any of the measures that have been discussed today. So if you are a member of the public and you wish to make a comment for the SMP, please take yourself off mute, raise your hand, and we'll call on you accordingly.

One last call. This is an opportunity for public comment. If any member of the public wishes to provide any comments for the SMP based on the measures that have been evaluated today, now is the opportunity to do so. Last call.

Okay. Seeing no hands raised and not hearing anyone taking themselves off mute, we will go ahead and move to the next step. Gabby and Hannah, I will give it to you.

## 0716e ePC-06 Unexpected Newborn Complications in Term Newborns

## Next Steps - Hannah Ingber

Ms. Ingber: Yes. Thank you. Thanks, Gabby. Yes, we'll go through the next steps quickly on the next slide.

So tomorrow' meeting will be from 1:00 p.m. to 3:00 p.m. Eastern Time. And our agenda has shifted a little bit. We'll discuss 0716e first and then have our measure methodology discussion for 2820 and 3687 -- oh, I'm sorry. We'll be discussing 3687e first and then 2820. That was also switched. So, again, 0716e, 3687e and then 2820e.

So that's our agenda for tomorrow. And I'll pass it to Christie and then Dave for any closing remarks.

Dr. Pickering: Christie, are you there? Maybe we'll go to Dave.

Co-Chair Nerenz: I thought I was waiting for Christie. We lost her. There's not much novel or unpredictable to say at this point. Thanks to everyone for all the diligence, the thoughtful comments and respectful treatment during the day. It's a long day. It takes a lot of attention. And a lot of good work I thought today. And we'll see if we can continue that tomorrow. Thanks, everyone, and have a good evening.

Meeting Adjourned

Dr. Pickering: Thank you, Dave. And last call for Christie. And we may have to look through the attendance list. I think we may have lost here. Okay. Okay. All right.

Well, I will also echo my thanks and appreciation to the SMP for all the work that you've done today and leading up this meeting. We still have a little bit more to go for tomorrow. So looking forward to reconvening with you all again.

So that's tomorrow afternoon on the Eastern side so we'll kick off right around 1 o'clock and then we'll get into our first measure, which will be 0716e after we do roll call.

So with that, thank you to the NQF team as well as our developers for all of the work that they've done for this meeting. And we'll see you all tomorrow. Have a great evening.

(Whereupon, the above-entitled matter went off the record at 4:30 p.m.)