NATIONAL QUALITY FORUM

+ + + + +

SCIENTIFIC METHODS PANEL
SPRING 2021 MEASURE EVALUATION MEETING

+ + + + +

TUESDAY
MARCH 30, 2021

+ + + + +

The Panel met via Videoconference, at 11:01 a.m. EST, Christie Teigland and David Nerenz, Co-Chairs, presiding.

PRESENT:

CHRISTIE TEIGLAND, PhD, Co-Chair
DAVID NERENZ, PhD, Co-Chair
J. MATT AUSTIN, PhD, Armstrong Institute for Patient Safety and Quality, Johns Hopkins Medicine
BIJAN BORAH, MSc, PhD, Mayo Clinic
JOHN BOTT, MBA, MSSW, Consumer Reports
LACY FABIAN, PhD, The MITRE Corporation
MARYBETH FARQUHAR, PhD, MSN, RN, American Urological Association
JEFFREY GEPPERT, EdM, JD, Battelle Memorial Institute
LAURENT GLANCE, MD, University of Rochester School of Medicine and Dentistry
JOSEPH HYDER, MD, PhD, Mayo Clinic
SHERRIE KAPLAN, PhD, MPH, UC Irvine School of Medicine
JOSEPH KUNISCH, PhD, RN-BC, CPHQ, Memorial Hermann Health System
ZHENQIU LIN, PhD, Yale-New Haven Hospital
JACK NEEDLEMAN, PhD, University of California Los Angeles
EUGENE NUCCIO, PhD, University of Colorado,

Anschutz Medical Campus

SEAN O'BRIEN, PhD, Duke University Medical Center
JENNIFER PERLOFF, PhD, Institute of Healthcare Systems, Brandeis University
PATRICK ROMANO, MD, MPH, FACP, FAAP, University of California Davis
SAM SIMON, PhD, Mathematica Policy Research
ALEX SOX-HARRIS, PhD, MS, Department of Surgery, Stanford University
RONALD WALTERS, MD, MBA, MHA, MS, University of Texas MD Anderson Cancer Center
TERRI WARHOLAK, PhD, RPh, University of Arizona, College of Pharmacy
ERIC WEINHANDL, PhD, MS, Fresenius Medical Care North America
SUSAN WHITE, PhD, RHIA, CHDA, The James Cancer Hospital at The Ohio State University Wexner Medical Center

NQF STAFF:

CAITLIN FLOUTON, MS, Senior Analyst
HANNAH INGBER, MPH, Senior Analyst
SAI MA, PhD, Managing Director/Senior Technical Expert
MATTHEW PICKERING, PharmD, Senior Director, Quality Measurement
CHRIS QUERAM, Interim President and CEO
SAM STOLPE, PharmD, MPH, Senior Director, Quality Management
SHERI WINSPER, Senior Vice President, Quality Measurement

ALSO PRESENT:

ALIXE BONARDI, Human Services Research Institute
KAREN FIELDS, MD, Moffitt Cancer Center
JACKIE GRADY, MS, Center for Outcomes Research and Evaluation, Yale School of Medicine
HENAN LI, PhD, Human Services Research Institute
KASIA LIPSKA, MD, MHS, BS, Center for Outcomes Research and Evaluation, Yale School of Medicine
KRISTEN McNIFF, MPH, KM Healthcare Consulting
JAMES POYER, Centers for Medicare & Medicaid Services
LISA SUTER, MD, Center for Outcomes Research and Evaluation, Yale School of Medicine
HUIHUI YU, PhD, Center for Outcomes Research and Evaluation, Yale School of Medicine

C-O-N-T-E-N-T-S

P-R-O-C-E-E-D-I-N-G-S

(11:01 a.m.)

DR. MA: Good morning, everyone. Welcome to the NQF's Scientific Methods Panel Spring 2021 Evaluation meeting. My name is Sai Ma. I am Managing Director and the Senior Technical Expert at the National Quality Forum.

I want to thank everyone for joining us today and want to welcome our SMP members as well as developers and the public who joined this meeting.

We really hope by using this virtual meeting platform we can try the best to improve interaction and information sharing. Feel free to use chat function or raise your hand.

First, I would like to express our appreciation for the time and the great work that our SMP members have put into reviewing a large amount of measures in this cycle, and we really look forward to a robust discussion today and tomorrow.

Second, I also want to thank the developers who submitted very comprehensive responses to the preliminary review within a very short window.

Those responses are included in the discussion guide which is attached to the appointment as Appendix A. So SMP members should have already reviewed those responses before today and tomorrow's discussion, and they will take into account of what was said in the responses when they lead the discussion.

And I would like to remind the developers as well when you provide a verbal response during the meetings, you don't have to repeat what you already put in writing, just to respond to what was discussed during the meeting.

And finally I wanted to say that even though SMP members volunteer their time for conducting those reviews and for participating in those meetings I do want to acknowledge that NQF staff's time, this work, and all the logistics are funded by CMS.

We would like to thank CMS for their financial support and for respecting our independence in the consensus development process.

At this point I would like to invite our Interim CEO, Chris Queram, and the Senior VP, Sheri Winsper, to offer welcoming remarks. Chris?

MR. QUERAM: Thank you, Sai. Good morning, everyone. I will be very brief. Last week I had an opportunity along with Sheri to meet with Sai and the NQF team, Caitlin and Hannah, to go over the materials in preparation for today's meeting.

It served to reinforce and help me develop an even deeper appreciation for the critical role that the SMP plays in upholding the scientific and clinical integrity of the measures that NQF brings forward for the field's consideration and use in all of the various forms for which measures are intended.

It also helped me appreciate the significant amount of time and effort that all of you invest in preparing for this meeting as well as the other meetings of the SMP during the course of this calendar year.

It's not a light undertaking, and I just wanted to express sincere appreciation to all of you for giving so generously of your time, energy, and expertise.

A special note of thanks to Dave and Christie for leading you through the next two days. And also, as I noted, a special note of thanks to all of you and to the NQF staff for the preparations that went into the meeting today.

So I wish you the best for the rest of today and tomorrow, and I look forward to observing and interacting with you as the meeting unfolds. Sheri, I turn it over to you.

MS. WINSPER: Thank you, Chris. I don't have a lot to add except for an additional just note of appreciation for all of you.

The SMP, Scientific Methods Panel, just contributes such a specific value to the work of NQF, the work of the partnership with our measure developers in informing them on specific technical expertise aspects of measure development in addition to really helping and serving as the key expert committee to our consensus development committees when it comes to looking at our complex measures and the scientific acceptability with reliability and validity.

I also will echo Chris's thanks. As we know there were a lot of measures to review this time and we know that this takes many hours of your time voluntarily and we so appreciate that as we couldn't do this without you.

I also want to thank the team as well, the NQF team, for all the work that they have put into preparing for this and thank you, Sai, for your leadership.

I look forward to a wonderful and rich discussion today and learning from you all as I listen. So have a great meeting.

DR. MA: Thank you, Chris and Sheri. I know Sheri really enjoyed all those scientific discussions. Now I would like to invite the SMP Co-Chairs, Dave Nerenz and Christie Teigland, to provide opening remarks and kick off today's meeting.

CHAIR NERENZ: Sure. Thanks, Sai. Thanks, Chris. Thanks, Sheri. And thanks, everybody. I won't --

DR. MA: Oh.

CHAIR NERENZ: I am told I had been muted. Am I back?

DR. MA: You're back.

CHAIR NERENZ: Okay. I didn't think I said anything that bad to start with. I'm told got muted. That's bad. I will repeat the thanks to everyone.

I just want to observe as we go

forward we have a very busy two days and just to observe, you know, what we do here is hard, it's complicated.

The statistical issues are complex. We have subtleties, we have nuances, we have slight variations in how things are done, and everything we talk about for the next two days are the close calls.

The ones that were obvious yes are now behind us and we are not talking about them and if there were any that were obviously no we're not talking about them either.

The ones we have in front of us are the hardest, the closest calls, is it low, is it moderate, and we just are going to spend two days sort of driving that territory, so thanks everyone for the diligent work.

I want to appreciate and thanks to the developers, the responses, at least the ones I looked at carefully, and I looked at many of them, most of them are on point, they are detailed, they are directly addressing issues that we raised in the preliminary, and that's good.

So we'll just try to work through that efficiently and make the best judgements we can.

CHAIR TEIGLAND: Yes, thanks, everyone. Also, I know one member of the SMP was traumatized, you know, by this round of reviews and needed a recovery period.

You know, this was a tough round of reviews. There were lots of measures that are really complex. It was interesting that Groups 1 and 2 each had a group with similar measures but, you know, one was sort of on readmissions, one was more on class, and, you know, the one set of measures didn't do well.

We reviewed them carefully. It really isn't that, you know, Subgroups 1 was evil, as David said, the people, you know, Subgroup 1.

The luck of the draw, right, and it just goes to show how very complex these issues are that there are nuances in even similar measures looking at similar, you know, outcomes that are very nuanced and really need, you know, careful consideration.

We are going to run into a whole slew of new issues, believe it or not, this time through and, you know, the calls are tough ones, you know.

We always want to try to balance between what the guidance says and what we know as experts, you know, our gut tells us about these measures, but to keep within the guidance and if we don't agree with it to change it or try to change it as we move forward to make this a better process.

So I think we need to keep that in mind as we go through today and tomorrow as we grapple with some of these issues and what goes into what bucket, you know, can we make this call based on the guidance we have or is this an area where, yes, this is the problem and we need to think about maybe making some changes to the guidance.

So those are some issues we'll be struggling with over the next couple of days. We've got a lot of measures to get through and a lot of issues, so let's gear up for a couple of good days.

I wish we were all in person, you know. Hopefully maybe next time. And we'll see where we land.

DR. MA: Great. Thank you. Thank you, Dave and Christie. I also want to thank you for your leadership and the actual work you put in behind the scenes to help us put together this meeting.

Very quickly a few housekeeping reminders before we dive in. We will have two meeting breaks. After the lunch break we will start reviewing measures and there will be another short afternoon break.

I want to remind everyone that the

meeting quorum is critical. For this cycle that means each subgroup will need to have at least six members today and tomorrow at the meeting to meet the voting quorum.

We do ask you if you need to step away, we understand that this is a busy time for everyone, if you do need to step away for any reason please let us know so we can keep track of quorum.

If we cannot meet a quorum for any reason for any part of the measure discussion and the re-vote we will have to vote offline and if we do have a situation like that we can talk about what the process looks like.

Feel free to use the chat features and the raising hand features, and we will try to take questions in order so every SMP member has a chance to speak up.

I would encourage everyone to mute yourself and only unmute yourself when you need to speak up. If possible do not use speakerphone. The audio quality is not as good as if you use the computer audio.

We also want to remind you that this meeting is being recorded and we have a court reporter to put together a transcript for this meeting.

So it is really, really important for you to introduce yourself before you speak up so we can write it down who can take the credit for the great ideas we hear on the phone.

Finally, for any technical support please send a chat to Hannah Ingber, send a message in the chatbox to everyone with some instruction there. If you have other questions regarding anything else about this meeting logistics you can also reach out to Hannah.

And finally, the reminder that our team sent you a link yesterday in the email for voting. If you do not have that link with you please reach out to Hannah now.

During the afternoon session and

tomorrow's meetings after we discuss each measure, a re-vote will need to happen and the link is where -- it's to the webpage where we are going to do the vote. So if you don't have that link now, please check your email or reach out to Hannah right now.

All right. I want to thank my team again. I think by now you should, probably are very familiar with our team. I do want to thank everyone on the team for their hard work and being really flexible and accommodating every request from the SMP members and from the developers.

For the interest of time I am just going to introduce my team. Michael DiVecchia is our Senior Project Manager. Hannah Ingber and Caitlin Flouton are our Senior Analysts.

Again, I also want to say that one more time that when you reach out to our team please use our -- please copy our mailbox, and if you want to reach out to any individuals listed here please do, but copying the mailbox will make sure, you know, your question is captured properly in case anyone of us is out of the office.

All right. So before we go to the roll call, I would like to remind everyone of our disclosure policy. Before the meeting each SMP member received a disclosure of interest form from us, and that is how we decided the assignment for subgroups, which is mostly based on everyone's conflicts of interest.

In the interest of transparency today we will ask you to orally disclose any information you provided in that form that you believe is relevant to the discussions today and tomorrow, especially specific to any of those measures being discussed.

I also want to remind everyone that remember you sit on this group as an individual. You do not represent the interests of your employer or anyone who might have nominated you to this panel.

We are interested in disclosures of both paid and unpaid activities relevant to the work today and tomorrow. At any point of the meeting if you realize that you or anyone else on the panel may have a conflict of interest, please let us know so we can address it in real time.

Okay. Without further ado, I am going to pass to Hannah to take attendance and ask her for DOI.

MS. INGBER: Okay, great. Thank you, Sai. I will, again, yes, ask you to unmute yourself and let us know if you have an disclosures to announce. Dave Nerenz?

CHAIR NERENZ: Yes. Dave Nerenz, Henry Ford Health System. I have been a consultant on Measure 0500, so I have recused myself from any involvement with that one. No other disclosures.

MS. INGBER: Thank you. Christie Teigland?

CHAIR TEIGLAND: I have no

disclosures.

MS. INGBER: Matt Austin?

MEMBER AUSTIN: Yes. Good morning. My only disclosure is I am part of the Measurement Development Team from Johns Hopkins that worked on Measure 3614, which will be discussed at the beginning of tomorrow's session, which is the stroke misdiagnosis measure.

MS. INGBER: Thank you. Bijan Borah?

MEMBER BORAH: Hi. No disclosures for any of the measures that will be disclosed both today and tomorrow.

MS. INGBER: Thank you. John Bott?

MEMBER BOTT: Hi. I was on a CMS TEP that gave council advice on 3501(e), but I am not on the team that reviewed that measure. So that's it. Thanks.

MS. INGBER: Thank you. And Daniel Deutscher?

(No audible response.)

MS. INGBER: Lacy Fabian?

MEMBER FABIAN: I am here. No additional disclosures for the measures within my group. Thanks.

MS. INGBER: Thank you. Marybeth Farquhar?

MEMBER FARQUHAR: I'm here. And I have no disclosures.

MS. INGBER: Thanks. Jeff Geppert?

MEMBER GEPPERT: Hi. Good morning. No disclosures today.

MS. INGBER: Larry Glance?

MEMBER GLANCE: Good morning. I have no disclosures. Thank you.

MS. INGBER: Joe Hyder?

MEMBER HYDER: Good morning. I have no disclosures.

MS. INGBER: Thank you. Sherrie Kaplan?

MEMBER KAPLAN: Since the last meeting or the last time we filled out the disclosures, I was appointed to the Technical

Advisory Panel for the Outpatient Pro-PM for the Yale Team, the CORE Team.

It's not related to the measures we are reviewing today, but I am not sure what that means in terms of recusal.

MS. INGBER: Okay. Thank you for announcing that. Thank you. Joe Kunisch?

MEMBER KUNISCH: Good morning. I have no disclosures.

MS. INGBER: Paul Kurlansky?

(No audible response.)

MS. INGBER: Zhenqiu Lin?

MEMBER LIN: Yes, hi. I think the measure for Yale CORE for CMS, so I will be recusing myself from discussing those measures.

MS. INGBER: Thank you. Jack Needleman?

MEMBER NEEDLEMAN: Good morning. No disclosures.

MS. INGBER: Gene Nuccio?

MEMBER NUCCIO: Good morning. No

disclosures here.

MS. INGBER: Sean O'Brien?

MEMBER O'BRIEN: Good morning. No disclosures for measures being discussed on this meeting.

MS. INGBER: Jen Perloff?

MEMBER PERLOFF: Hi, I'm here. No disclosures, but I am going to look for one for next time because I think it's pretty cool.

MS. INGBER: Patrick Romano?

MEMBER ROMANO: Good morning. I am here. I am recused on a measure that is not up for discussion this morning.

I will also just briefly mention in passing if my name appears in the Yale CORE measures of the EDAC measures for excess days in acute care, apparently I was involved in a phone call or two about 10 years ago when they were considering some of the original methodological questions behind those measures, but I haven't been involved at all since then, so we have determined that I am not recused.

MS. INGBER: Thank you. Sam Simon?

MEMBER SIMON: Good morning. No disclosures for the measures to be discussed.

MS. INGBER: Alex Sox-Harris?

MEMBER SOX-HARRIS: Good morning. No disclosures today.

MS. INGBER: Ron Walters?

MEMBER WALTERS: Hi. 3188, the 30-day readmissions for cancer patients, which I did not review, I was in the original development of it so I will recuse myself. I haven't been involved for a couple years now though.

MS. INGBER: Okay. Terri Warholak?

MEMBER WARHOLAK: No additional disclosures.

MS. INGBER: Eric Weinhandl?

MEMBER WEINHANDL: No disclosures.

MS. INGBER: And Susan White?

MEMBER WHITE: Hi. Good morning. I just have a disclosure for 3188, the 30-day readmission for cancer also. I wasn't on the review group, but it will come up for discussion, so thank you.

MS. INGBER: Thank you very much, everyone. I will hand it back to Sai now.

DR. MA: Thank you, Hannah. I will go over today's agenda really quickly. Just a reminder that the actual agenda is attached to the meeting appointment, so you can take a look at when the measures will be discussed and at what time.

Next, Caitlin is going to provide you a quick evaluation update for the Fall 2020 cycle and a quick overview of this current cycle. I will go over the process overview quickly and at a very high level some evaluation reminders.

We will take a break at noon and then after we come back we will start to dive into the measure discussions, and there will be a short break in the afternoon.

We will provide an opportunity for the

public to comment around 3:30-ish and then we will wrap up for today.

I do want to remind everyone we have provided a lot of materials in appointment again. The discussion guide is attached. That includes all of the measures that are submitted and slated for the SMP review for this cycle.

There are 29 measures that were deemed as complex enough that the SMP members should review them. So all of the measures, a brief description, and the SMP's preliminary analysis summary has been put into this guidance in the discussion guide.

In the appendix the developer's responses were included as well. So this is the document we are going to use throughout the day to help guide our discussion.

Also helpful to you I think are the three materials we put here on the slide. If you click on the link it will take you to the document. So those are the evaluation guidance from various points that NQF has developed.

I would say that the 2019 NQF Measure Evaluation Criteria and Guidance is the most comprehensive document for every measure type and new maintenance measure, and also a lot of the SMP review policy and process is included as well. If you are looking for any detail of the guidance, that is the document for you.

The last one is a very high level SMP Measure Evaluation Guidance. It's just a few pages long. It's, again, at a very high level. If you need a quick check that's a great cheat sheet for you.

All right. So based on the feedback we have received in the past that the SMP members really would like to have some kind of feedback loop built into the cycles because after you review measures, you don't necessarily know what happened to them at the standing committees and at the CSAC, so we are going to try our best at each meeting we will give you some update to each cycle.

So at this time I would invite Caitlin to provide a quick overview of what happened to the Fall 2020 evaluation.

MS. FLOUTON: Yes. Thank you, Sai. So at this time I would like to briefly walk through the Fall 2020 measure review cycle.

The SMP evaluated 25 complex measures that were submitted to this cycle, Fall 2020. Eight of those were discussed at our meeting this past October, and upon conclusion of the SMP's portion of this review cycle 20 measures passed both scientific acceptability criteria and continued on to the Standing Committee.

After further discussion the SMP did not reach a consensus on two measures. And those were also sent to the Standing Committee. Two measures did not pass scientific acceptability criteria, and they also did not get pulled by the Standing Committee for further discussion.

There was one measures that the

steward and developer decided to withdraw from consideration after reviewing the SMP's preliminary responses.

And then, of course, among the 22 measures that went on standing committees, two were re-voted on by the committee, that is an option that is available to them.

And so one of those re-votes resulted in the same passing rating that the SMP gave, but the other is listed here, Measure 0505, went to the All-Cause Admissions and Readmissions Standing Committee without a consensus reached by the Methods Panel regarding the measure's reliability.

After the Standing Committee discussed and voted on this measure, it did pass on reliability and is now recommended for endorsement waiting to be reviewed by the Consensus Standards Approval Committee, and that CSAC meeting is scheduled for June to be considered for endorsement.

Now going beyond this past cycle, here we have a table summarizing the outcomes of complex measures really since the SMP began in 2017.

So you will notice the number of measures the SMP reviews each cycle is variable, and this depends on what measures are submitted to us.

We also display here how many of those complex measures in each cycle passed and did not pass the SMP's review of scientific acceptability, with a note here also to acknowledge that data from this cycle, Spring 2021, is only preliminary at this time.

In nearly each past cycle there have been a handful of measures the SMP does not reach a consensus on, but those measures do continue on to the Standing Committee for their evaluation.

And then lastly we also look at how often the Scientific Methods Panel and standing committees agree upon the scientific acceptability ratings, with values that range from 72 percent up to 100 percent agreement.

Now I would like to switch gears to Spring 2021, our current cycle. For this cycle we had 29 complex measures submitted and assigned to the Methods Panel, of which nine were new measures.

We divided our SMP into three subgroups of eight to nine members and assigned each group nine to ten measures to review. And after those preliminary reviews 19 measures passed both reliability and validity.

There were seven cases where consensus was not reached. Five measures did not pass validity, and two measures were withdrawn after the SMP's preliminary review.

Thirteen measures are slated for discussion during this two-day meeting. We listed out also the measure types of all the 29 measures reviewed this cycle, most of which are outcome or cost/resource use measures, but we also saw three composites, two intermediate clinical outcomes, two PRO-PMs, two process, and one structural measure.

And then on this slide we provide a list of measures slated for discussion during this meeting and the order that we will discussing them, so starting today with measures evaluated by Subgroup 1, five of which are admissions and readmissions measures, and then we have one patient experience and function measure that we will be discussing later today.

When we reconvene tomorrow we will start with a neurology measure that was reviewed by Subgroup 2, move on to patient safety measures reviewed by Subgroup 3, and then a discussion of two renal measures evaluated by Subgroup 1.

And I would like to pause here to see if there are any questions about anything that we have just discussed.

MEMBER ROMANO: I have a question.

MS. FLOUTON: Sure.

MEMBER ROMANO: So this might be more of a question for the Chairs and perhaps Sai, but I wonder in these cases where there is a disagreement between the Scientific Methods Panel and the Standing Committee, of course, it's the Standing Committee's prerogative to disagree with the Scientific Methods Panel over one of these methodological considerations related to reliability or validity, but I wonder if is there any process by which our position is represented through subsequent discussion?

In other words, we're working very hard to try to take a consistent approach across measures and across different types of measures, and of course, the reason we are doing that is in part because the standing committees have not been able to do that in the past because they are focused on their particular domains of clinical expertise, which is very important, but it is our role to try to provide that methodological consistency across NQF's entire portfolio.

So as these measures go through the process, for example, and go to CSAC is there any representation of our interests as it were or of our reason for voting as we did which may differ from the Standing Committee's vote?

DR. MA: Yes. First, was that Patrick? Who was --

MEMBER ROMANO: Yes.

DR. MA: Okay. Good question. I want you to be able to take the credit for asking this important question.

We do share the SMP preliminary analysis meeting summary, so basically everything that has been discussed or reviewed, documented, is shared with the Standing Committee so they know exactly what your rationale is behind the voting.

However, NQF process policy is SMP members make this recommendation so the rating is not final or binding. You are voting on the reliability and the validity. It's a recommendation to the Standing Committee and that they could re-vote on reliability and the validity if they feel differently. So I just wanted to say that clearly.

MEMBER ROMANO: Yes, I understand that. I guess my question is the CSAC role to some extent is to ensure consistency across the entire portfolio.

And so does CSAC consider when there is a difference of opinion between the SMP and the Standing Committee on specifically reliability or validity scoring?

DR. MA: They could. So again, the whole history of how a measure being reviewed and voted from SMP to Standing Committee and, you know, comments we receive from public commenting period, the entire history of the review is summarized for the CSAC members so they could review from the perspective of whether or not policy or evaluation criteria has been applied consistently.

MS. WINSPER: I will just add, Sai. This is Sheri with NQF. I will just add she is absolutely right. The only thing I think I would just add to that is the CSAC's role is certainly to review everything that Sai just mentioned, but it's to review it in the context of was our process followed and do they think either all stakeholders, measure developer, NQF staff and the way we ran the process, did the committee follow the right process in thinking about it, whatever the issue may be.

But if we feel like over time that there is something that needs to be adjusted in that process that would also be something that we would want the CSAC to weigh in on so that it may enable, I don't know, maybe we want to just change something, either a process or a criteria, they would also weigh in on that.

But their main role is to weigh in on the consistency and following the process that we have outlined. I hope that is a helpful little bit of addition.

DR. MA: And Jack Needleman has your hand raised.

MS. WINSPER: I think you're on mute, Jack.

MEMBER NEEDLEMAN: Yes. Thank you, Sai. Patrick's question had two things, one was about the standing committees and the other was about the CSAC, and I know nothing about the actual CSAC process, but I do know a fair number of us are sitting on various standing committees.

I don't know if all of the standing committees have some member of the Scientific Methods Committee, but my experience both when I was on the Cost and Resource Use Committee and now that I am on the Admissions and Readmissions Committee is frequently we will be asked for more information about what the discussion was in the Scientific Methods Committee.

I try to be reasonably neutral in presenting all sides of those arguments, discussions, not arguments, discussions, but so you get some of that floating through the overlapping memberships, Patrick.

DR. MA: Zhenqiu is next. Zhenqiu, if you are talking we can't hear you.

MEMBER LIN: Oh, sorry, I was on mute. So I have a question. David may remember that in the past couple meetings I think at one point we talked about whether we should treat reliability and validity somewhat differently in terms of, you know, the view from Standing Committee, maybe you will give them more sway in terms of validity. Is this still the case?

CHAIR NERENZ: Yes. I had thought of that partly in response to what Patrick started here. I think that is correct. So in the area of validity, although the boundary is not crystal clear, there are some parts of that discussion that fall naturally to us, you know.

Were the methods used to establish validity statistically correct? Did they do the right test? Did they take the right data? Did they trim the data the correct way?

All of the technical things I think naturally fall to us, but there is the point where, for example, if we are asked to pass judgement on a quantitative metric, like two measures are correlated and the correlation comes to a certain level, is that sufficient.

That's kind of a gray area. If you recall from the Scientific Method Panel's point we may have a point of view and eventually we are asked to vote on that, but the Standing Committee may have a different view of the same number.

And when we transfer just a little bit into risk adjustment we see it even a little more clearly where we are asked to looked at the methods by which risk adjustment was done, you know.

Were the statistical models done correctly? Were there uses perhaps of a development set and a validation set, technical issues?

We often express our opinions on this group about, you know, were the right measures included, but that sometimes is up to us but sometimes it's not up to us. Sometimes it is up to the clinical expertise on the Standing Committee to say yes or no the right variables were included.

So in both cases there is a gray area where, you know, both groups, our group and the Standing Committee, may express an opinion but it may not be the same opinion and it may not be a significant problem if it is not the same opinion.

So I hope that is responsive to Zhenqiu's question. I think it comes up to me in risk adjustment, it comes up in validity, a little less so in reliability just because those issues tend to be more purely statistical in nature, but it could be in all three.

DR. MA: Thank you, Dave. Very

quickly I want to add one clarification for this slide. So for measures that did not pass the preliminary analysis or consensus were not reached, those measures will be automatically slated for discussion during the evaluation meeting.

However, for measures that passed the both an SMP member can pull that measure for discussion if they think that there are remaining issues or overarching problems to be discussed.

So a couple measures listed on this slide for fall into that latter category. I also want to add one clarification here is for Measure 0500, after reviewing the developer's responses the SMP member no longer wants to discuss this measure, so tomorrow we are just going to skip that portion of discussion.

All right. I think we can move on to the next section for this morning. I will provide a very brief process overview and evaluation reminders at a high level.

So the rating scales are important. The SMP members using those scales to help them differentiate the ratings for high, for moderate, from low, and then insufficient.

For example, moderate is the highest rating that a measure can be eligible for if only data element testing at the patient level was presented or if only face validity was conducted.

So it is important that you know there is a very comprehensive algorithm behind the high, moderate, low, insufficient ratings, and you can find the algorithm on page 24 and 25 of the guidance.

Sherrie, do you have a question about this slide?

MEMBER KAPLAN: Yes. It has come up before, but there is no meaningful difference between high and moderate, and my concern is over-complicating the algorithm if they are not going to use that information.

And so it's kind of about time that we

stop trying to discriminate between high and moderate if NQF isn't going to use the information. It should be high, moderate, and then low and sufficient.

I've made this argument before, but I rest my case.

DR. MA: Thank you, Sherrie. I actually reviewed the meeting summary from a few cycles before, and I saw your comment at that time.

We are going to collect the feedback from the standing committees next time we conduct our advisory meeting to see if anyone is actually using the information because we do present the proportion of votings for high, moderate versus low and insufficient and to see if that kind of granular level information provides any additional benefit, and we are open to a discussion that if nobody is actually using that information maybe if somewhere in the process we can generate a little bit more efficiency going forward.

But I will stop here for a second and just want to hear if any of the SMP members has an opposite view.

(No audible response.)

DR. MA: Hearing none, okay. We can also include that as a question in our post-evaluation meeting survey to the SMP members and if there is an overwhelming support for a binary voting going forward, pass, no pass, we are open to that option as well.

MEMBER KAPLAN: Sorry to interrupt you, Sai, but that wasn't my recommendation, just the collapsing of the high and moderate categories.

Low and insufficient to me feel like different things because it could be if you got the right data you would give it a higher rating than low, so I would not favor recommending collapsing into a binary.

DR. MA: Okay. That's helpful to

know. So sticking to what we have now, just a quick reminder for everyone for quorum we need to have 66 percent of SMP members within each subgroup, so for this cycle that means six members needs to be at the meeting to have a re-vote.

And once we start the vote we will combine the high and the moderate ratings to the pass category. So if we get more than 60 percent yes votes of the quorum we will get a pass. The consensus not reached will fall into 40 to 60 percent, including 40 and 60 percent, and then if we get less than 40 percent yes votes of the quorum, that measure does not pass voting.

Okay. In the next few slides just a quick reminder of the criteria SMP members are using to review each measure, and those slides were shared with the SMP members ahead of the time so hopefully there is no ambiguity here, but for the public's interest we will go over those criteria again.

So again, a lot of the criteria vary by the measure type. For health outcomes, intermediate clinical outcomes, cost/resource use measures, structure, and process measures, NQF does not require testing at both encounter and accountability level for either reliability or validity.

And I know we put it for new measures here, but actually for maintenance measures that's the same case. We prefer both. And again, we have discussed this at length in the previous meetings.

We are going to change data element level test to a patient or encounter level test. We are going to change the performance score level test to accountability entity level just to be clear and consistent.

We prefer both, but we currently do not require both. However, that does impact the rating as we just said before if you only provide a patient encounter level, formerly known as data element level test and result, you can only get as high as a moderate in rating.

Only when you provide accountability entity level result and testing that result, you can get it to as high as the high rating.

Face validity for new measures are accepted. For maintenance measures, we strongly, strongly encourage empirical analysis for validity.

However, if there is any real rationale behind why a developer cannot provide empirical analysis for the validity testing at the maintenance, the SMP members can consider their argument.

Alex, you have your hand raised. Do you want to unmute yourself and ask a question?

MEMBER SOX-HARRIS: Yes. Thank you. And thanks for clarifying that NQF does not require either data element or entity level, or only requires data element level reliability and not entity level reliability at either new or maintenance.

I had somehow thought that maintenance measures needed to provide entity level reliability, but that is not the case apparently. I would like to just put that on our agenda for a future discussion, whether that's the way it should be.

My preference would be that at maintenance there should be some kind of entity level reliability required, but I now understand that that's not the case.

And I know we have a lot of measures to go through today, but I have a clarification question that will impact my voting, which is in the case that a developer provides both data element reliability and entity level reliability results, and the data element reliability is excellent -- sorry, the data element is excellent, but the entity level is poor, given that they weren't even required to provide the entity level reliability, should I vote that the reliability is acceptable or not?

DR. MA: So I don't know, Dave or Christie, if you want to take this question, but I would just say because they provide both so your start point is high.

Then you start to consider the results, the testing, and then you can take into account those factors and then downgrade the ratings, but because they offered accountability level you start from high and then you go down as opposed to if they only provide data element testing, your go to place is moderate and then you downgrade from there.

CHAIR NERENZ: Yes, Alex, this is a great question and I have had some of these same thoughts myself and maybe we're talking about some of the same measures this time.

I mean you need wisdom of Solomon to try to figure this one out and I don't think there is written guidance about how to do it. You know, obviously, concerned in my own mind, I try to be fair to the developer in the sense that if they have given us more than is required they shouldn't be absolutely failed if in the additional information there is something that looks weak.

If they have passed the basic requirement I sort of feel some obligation to pass it through. I certainly would not give it a high rating.

Now having said that though, in the specific example you gave what I struggle with is that when I think about what kind of reliability is most important to me I always look at the entity level as being fundamentally the most important and then there comes the tension and I have to go back and say, you know, if I am faced with this what would I do. I don't know.

It's an absolute quandary. You know, what I might have done is rate it moderate but then put in the notes to say essentially a summary of what you said, at the data element level things are absolutely fine and that's what they are strictly required to provide.

At the entity level though where I think it's more important it's not good, in fact, it's not even acceptable, and I just would let the narrative try to carry the message, but this is really tough.

MEMBER SOX-HARRIS: Yes, and we have several measures where this is the case where we have incredibly good item level reliability and in my view quite poor score level reliability or entity level reliability, and I think the reason why we had -- we're either passing measures with bad score reliability where there is consensus not reached is this technical problem which I think could be solved if the entity level criteria vote is required in the new level reliability and explicitly preference it for maintenance measures. That's all.

DR. MA: Thank you, Alex. We will definitely bring this back to the advisory meeting in May, and we can talk about how to update our evaluation guidance.

Some quick reminders about composite measures. Composite measures do have a few additional requirements. Those measures, their components of the composite measures should have their own properties of reliability and validity.

And we have mentioned this throughout the last few advisory meetings, but multi-item scales of survey or questionnaires, like of CAHPS measures, are not composite measures by NQF's definition.

We do require reliability testing of the composite measures at the score level at the accountability entity level, so that is a different measure type from the previous slide.

The developers can show reliability testing of the components level, but that is not sufficient. Score level validity, however, is not required until maintenance.

And as Sherrie has talked to us a

couple of times during our past meetings that for composite measures it's really critical to talk about the measurement model, assume that the developer either is a formative model, a reflective model, and how they construct that the composite is the key for the SMP members to assess whether the analytical strategy for testing reliability and the validity is appropriate. So those are some things to keep in mind.

Instrument-based measures and survey measures, again, they are different. For those measures for reliability and validity testing is required at both levels.

Some general reminders. Testing must align with specifications. For example, if the measure includes multiple levels, including practice level and hospital level, then each level needs to be tested separately.

If a measure is checked for multiple levels but they only provide tests at one level, normally NQF staff will capture this inconsistency in our initial review and the triage, and we will contact the developers right away.

But if for whatever reason that happens during the review, the SMP members could pass part of the measure for the level they think is appropriate.

Occasionally there are several performance measures included under one NQF number. Those measures need to be evaluated separately, and you can pass a part of those -- you can pass some of the measures, not every measure.

So for example, this cycle there was a CAHPS measure, but it -- there is one NQF number for a CAHPS survey, but it included 17 different measures, so you could pass a portion of those measures but not the rest based on your review.

About risk adjustment, inclusion or not of certain risk factors in the model should not be a reason for rejecting a measure.

However, if you have strong concerns about the discrimination, calibration, or overall method for adjustment those are grounds for rejecting a measure.

For all measures incomplete or ambiguous specifications are grounds for rejecting a measure. However, you know, we did offer an opportunity for the developers to provide clarification so please take that into account.

Empirical validity testing is expected at time of maintenance. However, if for some reason a justification is provided the SMP members can review the justification.

Patrick, you have your hand raised. You can unmute yourself.

MEMBER ROMANO: Yes. Yes, I was wondering on the previous slide if you could go back to the last bullet point.

So we do have an example here today, Number 3622, National Core Indicators for

Intellectual and Developmental Disabilities, Home- and Community-Based Services, this does appear to include a number of performance measures.

So how would the voting be conducted on this if we believe that some of the components meet criteria but others don't?

DR. MA: That's a good question. After the discussion, we can vote each measure.

MEMBER ROMANO: Because that could get a bit tedious because of the number of measures. I believe there are 14 measures in total within that NQF number.

So it's just -- and we saw this also for the HCAHPS measure that you described. So it is rather difficult and burdensome, you know, to go through.

Clearly the component, you know, these 14 measures are heterogenous, and so it's difficult, you know, to pass them all and I guess that we should put on the agenda for discussion separately, you know, some recommendations for developers because packaging 14 measures under one number really becomes awkward when they are so heterogenous.

DR. MA: Yes. Gene?

MEMBER NUCCIO: Yes, real quick. If you could go back a couple slides to the composite. Just a clarification, NQF in terms of functional behavior measures has measures that evaluate a patient's ability to bathe, ability to dress upper and lower body, to ambulate, and those are four distinctive measures already approved by NQF.

And as I understand composite measures, if the developer was going to combine those particular measures into a single patient functioning measure, then that would be considered a composite measure.

However, suppose the developer instead re-conceptualizes how they wanted to evaluate bathing, upper/lower body dressing, and ambulation differently from the way those individual measures are but then combine them into a single functional measure.

So they are not using the existing measures, they are creating new measures, but combining very different kinds of behaviors, all of which deal with patient functioning. Is that a composite measure?

DR. MA: So I don't think it's -- my NQF colleagues, feel free to jump in and other SMP members feel free to jump in, but my understanding is we do not require the components to be NQF-endorsed measures.

As long as they are measures, the components have their own reliability and validity, then you constructed them to a composite measure, that still is a composite measure. And I see Dave nodding, so thank you.

MEMBER NUCCIO: Okay.

DR. MA: We are over time. I am going to entertain one more question from Larry.

MEMBER GLANCE: Hi. Thank you. This is on the most recent slide where you talked about looking -- specifically looking at the factors that are included in a risk measure, and saying that inclusion or not inclusion of certain factors should not be a reason for rejecting a measure.

I just wanted to comment that the measures of discrimination and calibration are at times not terrifically granular in terms of being able to differentiate between a good risk adjustment model and a not-so-good risk adjustment model.

And I do believe that it is our obligation as a panel to look under the hood and look at which risk factors are included, and whether or not we believe that the list of risk factors -- not necessarily whether one or two risk factors are included or not included, but whether or not the risk factors that are included are reasonable.

So for example, if you have a surgical mortality model, and if it's applicable to a very heterogenous group of surgical procedures, one would expect to see some measure of surgical complexity.

And if you did not, then even if the model had say acceptable discrimination, acceptable calibration, I don't think that one would necessarily want to pass that model in terms of being appropriate risk adjustment.

So I think it's important that we don't just sort of kind of casually say look, you know, it's not about which risk factors are included or not included.

I think that this panel should actually look under the hood and should evaluate the content of the model, not just its performance, because it's possible when you are looking at a very heterogenous patient population to have a great C-statistic and still not really have a very good model.

DR. MA: Thank you, Larry, point well taken. I think we have a couple of other reminders to go over really quickly before we break out for lunch break.

So we have articulated in the guidance that for the SMP members to do their thorough review, there is a strong desire for more details when you describe the methodology of the testing.

We also require more than one overall statistic if reporting on signal-to-noise reliability. So distribution of the statistic is really useful.

And the desire for detail in description of construct validity narratives and all those information listed here are super helpful for the SMP members to be able to do a fair evaluation.

The lack of Number 2 and 3 should not be grounds for rejecting a measure. Number 1 is really important because if we don't understand the methodology, we don't understand what you did, how to interpret the results, then we cannot do a fair evaluation for this measure.

The last two slides, really quickly, on the process, what happens to a measure when it doesn't pass SMP. So for measures that do pass SMP review and where consensus cannot be reached, those measures will be moved forward to the standing committees and they will evaluate and make recommendations for the CSAC.

For measures that do not pass the SMP, if they are eligible they can be pulled by a standing committee member for further discussion and a possible re-vote.

In the next slide we are going to talk about what eligibility means here. So this decision aligns ways NQF staff and SMP co-chairs, so after the evaluation meeting, we will do a quick touch-base on whether or not at the end of the day for those did not pass measures, do they meet the eligibility criteria?

The bullets here are the scenarios

where a measure is not deemed as eligible for a re-vote by the Standing Committee. I think it kind of answered Patrick's earlier question.

So if inappropriate methodology or testing approach was applied to demonstrate the reliability or validity or the calculation is incorrect or formulas were wrong, description of testing approach, results, and data is insufficient for the SMP to do a fair and solid evaluation, or appropriate levels of testing are not provided or otherwise did not meet NQF's minimum evaluation requirements.

So if a did not pass measure falls into any one of those four scenarios after a discussion with the SMP co-chairs, we will make a recommendation to the Standing Committee that they should not pull such measure for further discussion, and that decision will be shared with the developers in real time.

All right. We are 10 minutes over. I do want to offer everyone 20 minutes break and take a quick bite as we have a really full agenda for the afternoon.

So I am going to -- so I will mute everyone for now, and please join back at 12:30.

(Whereupon, the above-entitled matter went off the record at 12:11 p.m. and resumed at 12:31 p.m.)

DR. MA: Hannah, do you want to do a quick roll call for Subgroup 1 before we get into the discussion?

MS. INGBER: Sure, I can do that.

DR. MA: As you call off the names, I just want to go over the process of the discussion. NQF staff will introduce each measure. Then the SMP assigned lead discussants will summarize the key concerns and that they will take into account the responses we have received.

Then other SMP members will be invited to comment and since we have a really packed agenda, I would advise you not to repeat what has been already said, only add additional comments at the time.

And then the developers will be given two to three minutes for initial response and again, for the developers, please also try to focus on what was discussed during the meeting and not to read the whole responses in writing.

Then the discussion will be opened up to the full panel, but recused members cannot participate in the discussion of voting.

Developers can then respond to those questions. Finally, we'll move on to a vote and only the subgroup members who reviewed are the member that can vote on this measure again.

I will pause and see if there is any question about the process.

All right, hearing none, Hannah, do you want to do a quick roll call for Subgroup 1 as all the readmission measures are reviewed by the first group.

MS. FLOUTON: Hi, Sai. This is

Caitlin. I will do the quick roll call. Hannah just needs to grab her charger.

So for Subgroup 1, if you don't mind just unmuting your cell phone and call your name just to acknowledge you are present, starting with Eric Weinhandl.

MEMBER WEINHANDL: I'm present. Sorry, on mute.

MS. FLOUTON: Thank you. Sean O'Brien.

MEMBER O'BRIEN: I'm here.

MS. FLOUTON: Great, thanks. Sherrie Kaplan.

MEMBER KAPLAN: I'm here.

MS. FLOUTON: Wonderful. John Bott.

(No response)

MS. FLOUTON: I will circle back. Larry Glance.

MEMBER GLANCE: Here.

MS. FLOUTON: Thank you. Terri Warholak.

MEMBER WARHOLAK: Here.

MS. FLOUTON: Thank you. Patrick Romano.

MEMBER ROMANO: Present.

MS. FLOUTON: Thank you. I'll check for Paul Kurlansky, as well, who may not be present.

(No response)

MS. FLOUTON: And Dave Nerenz.

CHAIR NERENZ: Here.

MS. FLOUTON: Thank you. I'm going to circle back to see if John Bott is on the line.

(No response.)

MS. FLOUTON: Okay. Is there anyone that I missed? I believe that's the group.

Okay.

DR. MA: All right, so we do have a quorum. Thank you, Caitlin.

At this time, I would invite our director Matt, to go through the readmissions measures.

Matt, do you want to introduce yourself first?

MR. PICKERING: Sure. Can you hear me okay? Great. So my name is Matt Peckering. I'm a senior director here at NQF. I'm working on the all-cause admissions and readmissions portfolio. It's a pleasure to speak with you all again as well on this meeting and also for tomorrow as well I'll be on the call.

For the readmissions measures, I'm going to start out with 2880 which is excess days in acute care after hospitalization for heart failure.

Are we going to be voting on this, Dr. Ma? Should I proceed, Sai?

DR. MA: Yes, thank you. Sorry about that. Caitlin, are you ready for a voting test?

MS. FLOUTON: Hannah, are you running the voting test?

MS. INGBER: Yes, I'm ready.

DR. MA: Okay, so before we dive into each measure discussion, this voting test is for the whole panel. Again, you should have the link and we are going to run a test just to make sure everyone has access and you can vote when the time is ready.

All right, Hannah, take it away.

MS. INGBER: Okay. Let me just set it up real quick and I'll let you know when it's active. Yes, again, if you need the link just message me. Just one minute.

Okay. You should now see for the methods panel members in Subgroup 1 a question saying a test, and your options are yes and no.

We're getting some responses in. Let me just check on those. Okay.

We have 13 responses for yes and 1 response for no. So people not in Subgroup 1 have voted, but all the members in Subgroup 1 are present.

DR. MA: All right. Thank you.

MS. INGBER: Thank you.

DR. MA: Now we can move on to the measure discussion. Matt?

MR. PICKERING: Yes, sorry about that. Thank you.

DR. MA: No, thank you.

MR. PICKERING: Again, it's a pleasure to speak with you all once again. So I'll be going over all cause of admissions and readmissions measures.

And the first one, as you can see on the slide, it's 2880, Excess Days in Acute Care after Hospitalization for Heart Failure.

The measure developer for this measure is Yale CORE and this is a maintenance measure and I'll just go through a brief description of the measure. This measure assesses days spent in acute care within 30 days of discharge from an inpatient hospitalization for heart failure to provide a patient-centered assessment of the post-discharge period.

So this measure is intended to capture the quality of care transitions provided to discharge patients so that a heart failure hospitalization by collectively measuring a set of adverse acute care outcomes that can occur post-discharge and those including emergency department visits, observation stays, and unplanned readmissions during the 30 day post-discharge.

So in order to aggregate all three events, the developer measures each in terms of days and the Centers for Medicare and Medicaid Services annually reports the measure for patients who are 65 years and older, are enrolled in Medicare fee for service, and hospitalized in non-federal short-term acute care hospitals. So this is an outcome measure. The data source is claims. There are other data sources used as well for aspects of exploring potential risk adjustment such as the Census data, American Community Survey, and the Veterans Health Affairs administrative data, Medicare enrollment data, et cetera.

Obviously, at the facility level, it

is risk adjusted with 37 risk factors. I won't spend too much time on reliability as the subgroup in their initial evaluation of this rated this as moderate for passing the measure with moderate reliability.

I'll just touch on validity before I turn it over to my colleague, Dave, to do a summary of some of the key concerns.

For validity, there was consensus not reached. The validity testing conducted was at the measure score level, but they also had face validity was assessed, using a survey based information provided from the 16 member technical expert panel which more than 80 percent of the experts moderately or strongly agreed with the validity of the measure.

Regarding the empirical validity testing, construct validity was assessed as the relationships between the heart failure of the measure, measure score, and risk standardized readmission rate. Group scores and the overall hospital rating score in the heart failure readmission measure.

The developer hypothesized the relationship between heart failure, excess days, measured scores, Star Rating readmission score group, and Star Rating summary scores. They also hypothesized a positive relationship between the heart failure and the excess days measured scores and the heart failure readmission rate score.

Part of the results from this, they were varied, but as hypothesized with those correlations. And regarding risk adjustments, the developer found a c-statistic of .59 and an R squared value of .027.

There were two social risk factors that were tested and were found to be statistically significant, but do not appear to meaningfully affect hospital performance estimates and were therefore not included in the risk adjustment model.

This information as well as was found

in the discussion guide which is on page 7 and 8 if you also want to look at this in more detail.

But Dave, I'll turn it over to you to maybe highlight some of the key (audio interference).

CHAIR NERENZ: Thanks. And Matt, actually, there's so much to add, this is beyond what I had planned to say at this point any way.

I thank our good folks at Yale CORE with a very thorough, very detailed response, and I think now we can all assume that the Subgroup 1 members have had chance to read that, think about it.

So I think what we should do at this point is focus the discussion, we're going to turn to the developer here in just a second, on the questions of validity, recognizing that the re-vote will be on reliability and validity. Reliability was a pretty clear path.

The two issues, well, there's really one main issue with validity that came up in a number of the comments that I think led to the pattern of votes we see here. The observation that the same readmission events were occurring in both sides in the correlation, so to speak, of what we use for validity and that is that there were readmissions that appeared in the acute data measure, but also they appear in a readmission because it's the same readmission. So the response clearly spoke to that.

And I think it would probably help us if the developers now could just focus on that, talk us through that, illustrate the high points of what you think the adjustment in analyses, if that's what we call it, or the revised analyses show and draw us attention to that.

I think secondarily we have the issue of risk adjustment in the c-statistic that's probably worth a little attention. But there are a few other things that raise a number of issues in the response that really don't need to be discussed here unless one member of the subgroup wishes to focus on that.

But why don't we start with -- show us the high points of the validity response and the c-statistic adjustment response and then we'll see where we go from there.

DR. MA: Larry's hand is raised. Do you want to jump in right now?

MEMBER GLANCE: Yes, thanks. So I think one of the things that I really think is foundational that was not addressed in terms of the validity discussion is the fact that when the new model and CMS, or rather the Yale group, updates their model on a periodic basis, so they go back and they take the model and they re-estimate it using newer data and then they use that to customize the model coefficients.

Now when they did that and they did their validity testing, so when they looked at model discrimination and model calibration, they did that using the entire data set that was used to re-estimate the model. So they didn't divide up the data into a training set and a validation data set as is customary. I think we all recognize that if you do not do that, then you end up with overly optimistic measures of model performance. And I think it's pretty much a standard that when you're evaluating model performance, you cannot evaluate model performance using the same dataset that was used to estimate the model.

This is -- I think is foundational. It is a, I believe, a really important issue with all three of these measures. And I find it a little perplexing that the measures when they were initially presented for endorsement were, in fact, developed and validated using independent datasets. And yet now when the measures are being presented to the NQF for re-endorsement, which one would think that there would be even a higher bar in terms of measure performance, that what is being presented now is that, "look, we're just going to give you the model. We're going to evaluate. We're going to test the model performance using the same data and that should be sufficient." And I really don't think that is sufficient. And I don't think it meets the standards for evaluation of model performance.

So I think this is an important point, not just for our subgroup, but for the entire panel to discuss and address.

DR. MA: All right. If no other subgroup member wants to cut in at this point, we will invite our developers to provide a response.

MS. GRADY: Hi. My name is Jackie Grady. I'm Associate Director at Yale CORE of the Data Management and Analytics group.

Thank you very much for your comments and I just really want to thank the whole entire Scientific Methods Panel and the staff the time they've taken to review these measures.

In response to your concern about the development and validation datasets and how we look at model performance. That's correct. For ongoing model performance we do not create a training and testing set. But I think that it's addressed in the presentation about what was done in the original development. They were not two separate datasets. We did have a development and validation set when the measures were first developed back in 2016 or earlier.

I think it's certainly something that we can take back and think about when we are revising our models each year, but as these measures are in performance, we do look at the basic model performance statistics with new data each year and as they are -- sorry -- as specifications are changed or if new coding is added, we need you to look at that.

I'm not sure if there's anyone else on our team that would like to respond at this time. That's all I have.

DR. SUTER: Hi. This is Lisa Suter from Yale CORE. Can you hear me? Am I unmuted?

DR. MA: Yes, we can hear you.

DR. SUTER: Great. Thank you, Sai. So just following up on Jackie's response, I think you had a clarification, so we're not reselecting risk variables every year. We're just recalculating beta coefficients for the risk model on a yearly basis. So every year that the measures are updated, there are three year rolling measures. The measures are updated with new data that allows us to reselect new risk variables.

And yes, and for this current evaluation to look at measure validity, overall measure validity, we're using the same data that we used to update the beta coefficients which represents the most up to date data for calculating the measure in order to understand for the same data period the performance of the measure compared to all the comparators which I think Jackie will respond to because that was the main concern -- our understanding in the comments was the main concern about the overlapping readmission events, which I believe we have addressed.

And we will certainly take back the feedback from the committee if the committee as a whole feels like this is critical for ongoing maintenance of the measure that the measure result validation work is separate from the most current updated beta coefficient for the risk adjustment model, but from our standpoint, we review all of the data in order to re-estimate the data coefficient, we would have to use less data to really update the information in order to divide that into a training and validation set which is very critical for this group.

And if the SMP gives us that feedback that we should do that in the future, we absolutely will. But I just wanted to make a distinction between what was done during the original measure development versus what we do for maintenance of the measure and maintenance of measure endorsement. Thanks very much.

DR. MA: Thank you. I just want to open the discussion to anyone on the SMP.

And Patrick, you had your hand raised.

MEMBER ROMANO: Sure, good morning. So the same issue, of course, applies to all three of these measures. So I'm wondering if the Yale team might be able to clarify a little bit. So to put it in context, all of these measures have relatively poor discrimination and we're used to that by now. All the readmission measures, all the EDAC measures are all hovering around .6 in terms of the c-statistics. And we realize this is a little bit more complicated because of the two-stage hurdle model and so with really now we're focusing on the logic part of the model.

But it obviously is important if you're already starting at .6 and that discrimination were to be substantially lower in a set-aside sample. So what you report here is the c-statistics from the development dataset of .507 and the c-statistics from the testing dataset which is the -- if I understand the entire updated dataset from 2016 to 2019 of .59.

Now at some point in the past, is it the case that you did assess the model on a set of five samples, it doesn't use in the development at all and if so, did you see a substantial decrement in the model discrimination when you did that several years ago?

MS. GRADY: We would have to pull up those numbers, but no, we do not. The original process for development of a measure was to take two years of data and randomly split that and create a development and a validation dataset. And we did run all the calibration -- sorry.

DR. SUTER: Sorry, Jackie. I didn't want to cut you off. I thought you had stopped.

MS. GRADY: No. I was just saying that we did actually run the statistics on both of those datasets. Go ahead.

DR. SUTER: Thanks. Our team is working on pulling that information, but the c-statistics were not dramatically different, the measure was re-endorsed, or endorsed at that time as were -- and the same measure methods were used by the readmission measures that were seen in the last cycle and moved forward by the SMP for committee review.

DR. MA: Thank you. I have Sean and then Sherrie.

MEMBER O'BRIEN: I guess I want to give a slightly alternative perspective. I think the lack of a split sample training and validation set is actually does not bother me and I've encountered this issue as a measure developer as well. I think this split sample analysis addresses the types of analyses that we as reviewers expect to see. I'm not sure they really address the true underlying questions that are important.

So in the context of case mix adjustment we want to know that all the important factors that may be associated with outcomes and differs systematically across the provider units are addressed in the risk model and if that model is correctly specified in the sense that it's making mathematical assumptions, then we need to know the assumptions are correct and valid in order to be able to be guaranteed lack of bias.

And from the standpoint of assessing presence of large misspecification errors, using internal sample for that purpose actually turns out to be adequate and potentially more powerful than a split sample approach.

When you're using a hierarchical model for case mix adjustment you're re-estimating those coefficients every time you estimate provider performance and you expect that type of model to overfit the data and if you evaluate performance in that internal test that you'll have exaggerated measures, for example, of discrimination or calibration compared to an external sample. But to the extent there's overfitting, an appropriate statistical approach will consider that overfitting to be part of the overall uncertainty in the measure and that uncertainty will be propagated into the calculation of confidence intervals, so it's a random error, rather than necessarily a systematic error.

And to the extent that it might lead you to over re-estimate the c-statistic, I guess I have a different take on that as well. I think that you have to evaluate model performance in the context of what you're trying to do and it's not necessarily necessary to go and accurately predict which individual patient is most likely to have an outcome and which individual patient has probability close to zero, it needs on average to be well calibrated to adjust for case mix differences. And you can have perfect ability to remove or minimize confounding from case mix by using a model that has a low c-statistic. They're not directly -- the c-statistics are not directly measuring a model's ability to account for bias. So the c-statistic's estimate may have been exaggerated, there's not actually a c-statistic number that would cause me to be concerned about the validity, but the model in isolation without some kind of comparison to some alternative model that adjusts for a different set of candidate predictors.

DR. MA: Thank you. Sherrie, can you unmute your phone?

MEMBER KAPLAN: Yes, thank you. I have just a sort of a little issue, but it's related across. At Table 1B where you show the relationship between the Star Rating standardized summary score excluding the entire readmission group, and it occurs for all three, 280, 81 and 82. It's all titled the relationship between pneumonia EDAC and the Star Ratings for that.

I'm assuming that each one of those Table Bs applies to the condition being studied, heart failure, pneumonia, and AMI, right?

MS. GRADY: Yes, that's correct. I'm sorry, that is a typo.

MEMBER KAPLAN: Okay. Then this is related to the issue of sort of how these validity variables relate to the sort of -- validity variables that you've chosen relate to each one of those measures. And that is, when you exclude the Star Ratings, the standardized readmission scores, you get a drop and it's a significant drop between the correlation coefficients not from zero when you exclude the specific individual conditions readmission scores. But then when you exclude the entire readmission score, you get a drop from explaining about 33 percent of the variance thereabouts to four percent of the variance.

So how does that figure into your validation story?

MS. GRADY: Sorry, I was muted. You know, I thought about doing this particular approach was really sort of mixed. We really didn't expect there to be a very strong correlation there. The Star Rating is set up so that there are different -- measuring different domains of quality. So the idea that they would necessarily be highly correlated to the remaining domains if we were to take out the readmission group scores part of it, we really didn't expect it to be a very strong correlation. We did present it and it is going in the correct direction and it was specifically significant, but we really didn't expect it to be a very strong correlation.

MEMBER KAPLAN: So this is a stretch for you, but it works from somebody who is observing these data across three different conditions. The overall score is mainly being driven by readmission. Is that part of this validation story or that's not something you want to even go?

MS. GRADY: It's not necessarily where we want to go with it. But like I said, we decided to just present that information for, you know, complete transparency about how it relates to the Star Ratings methodology in total.

MEMBER KAPLAN: Fair enough. Thank you.

DR. SUTER: Sorry, this is Lisa, just jumping in, Sherrie, for you. I think these measures were created because there was a concern of unintended consequences of the admission measure as not capturing all returns to hospital. I think that CMS is on the call today, but the goal in these measures are to provide supplementary information that, you know, the fact that there are significant correlations, I think, argues for these measures not being unimportant as a contributing information understanding readmissions and other returns to the hospital.

We do know for hospitals that do really well on reducing admissions, many of them do increase emergency room rates in order to accomplish that. I think that's a very complicated clinical response in terms of understanding what it is that allows you to bring someone into an emergency room, evaluate them, discharge them and we do not see increased adverse events like mortality associated with that. So the whole process is compressed, but we are looking at the additional validation results and we're very happy to modify our processes if the committee would like to see different or -- you know different approaches or can provide us guidance.

DR. YU: Hi, this is Huihui Yu from Yale CORE. We actually have considered two parts of the construct applicability, discriminate and convergent validity. So if the Star Rating, the overall scoring including the readmission group scores, we are expecting it to be correlated with EDAC measures.

However, as we include the readmission group scores, we don't expect the correlation should be strong. So we provide those to actually provide the evidence of both convergence validity and discriminate validity. Hopefully, this will be helpful. Thank you.

DR. MA: Thank you. We have time to entertain -- I was just going to say Larry gets the last question. Then I see Jack's hand's up. So can you discuss quickly of your concerns and we will go on our vote.

So Larry and then Jack, you can unmute yourselves.

MEMBER GLANCE: So I just want to push back a little bit more again about the need for separate datasets for model estimation and model validation. I completely agree with Sean is that the final model, when you customize it, should be based on the entire data set because the end of the day when you're looking at your P/E ratio, it is going to be based on all the data, not on half the data.

But what I am suggesting though is

that the same standard that we use for model validation for the original endorsement process should be still the same candidate we use when the model comes back up, when a measure comes back up for re-endorsement.

I'd like to add to the fact that this is a new model. When their model was first developed back in 2010, it was based on ICD-9 codes. It's now based on ICD-10 codes. There's not a one-for-one match between the ICD-9 codes and ICD-10 codes. This is really a different model.

And I think if we're going to look at model validity, we need to do it the way that I think all of us have always understood that model validation should occur, meaning that you have two separate datasets, one used to estimate the model and the other one used the validate the model.

And then, I agree with Sean, that the final model should be re-estimated using the entire dataset, but I don't think that it's sufficient to look at model performance using exactly the same data that's used to estimate what ends up being a really different model and looking at model performance using the same data set used to estimate the model.

So I honestly think it's an important issue and I'm a little bit surprised that no one on the panel thinks that this is something that they would want to discuss further.

DR. MA: Jack.

MEMBER NEEDLEMAN: I'm not going to follow up on Larry's thing. I do think it probably merits some attention, possibly in our monthly meeting, just so we can have a little bit more time and space to have consensus, build some consensus around that.

I have a question for the developers about the risk adjustment from a different perspective than the statistics.

Which is, I'm taking the statistics as

a given. You don't have a very discriminatory model for who's getting readmitted and who isn't. That's what all the statistics are saying.

And we've got measures which have, are specific to specific diagnosis.

So, we're eliminating some of the variability across, you know, rates of readmission, different across different diagnosis. So, we've eliminated some of the sources of variability that would be easy to capture in risk adjustment.

But, these low statistics seem to imply that the -- in general, either we've got pretty homogeneous populations in these groups. So, a risk adjustment model doesn't discriminate, because there's not a lot of variation to discriminate against.

Or, the things we do measure that have a variance, don't seem to affect readmission very much. And you've looked at this model. You've been looking at the data.

Is it just that we don't -- you know, what is going on with readmissions? Do we not -- are there not factors that really try to predict readmissions for the full ED, or observation days?

Do we have -- are these models inherently, once we go to specific diseases, going to have very low discrimination among patients? And very low explained variance?

MS. GRADY: So, just to address that, I mean, I think there's a couple of things. You know, we do feel that a lot of variation is really at the hospital level. Which we do try to capture with a hierarchical model.

Some of the other, the factors that we would love to include that we think would be highly predictive, such as functional status, are really difficult for us to put -- to have data for right now.

Those are something that I know CMS is looking into on how we would come -- finding a better way to do that. And also, you know, I think the risk adjustment part of these measures is really about trying to level the playing field across possibles, not necessarily to predict readmission.

So, all of that is in place. And I think we feel comfortable. Not comfortable, but that's the reality of what these models look like for readmission.

I you know, I'm wondering if Lisa has anything to add to that. If she would like to.

DR. SUTER: Thanks Jackie. This is Lisa. I'll just note that, you know, the needs measures, not the EDAC measures, but the original readmis -- 30 day unplanned readmission measures for myocardial infarction, heart failure, and pneumonia were originally developed, there was -- they were validated against clinical, you know, models using clinical data from registries to, you know, really to try and understand the relationship of ICU 9 codes to clinical predictors.

And, you know, just as an antidote, if you tried to mimic the clinical risks model that you could create from clinical data, you know, like ejection fractions, and you know, vial fronts and things like that, claims do really poorly.

And when we let the claims sort of function as their own data element, and we use the data to predict readmissions and mortality for these original six measures, not the EDAC, but the readmission mortality measures, we were able to predict much better. And commensurate with the clinical risk models.

So yeah, we had not done clinical validation of these models, but, you know there have been multiple ICD measures that have transitioned from ICD 9 to ICD 10 that have been in front of this measure and this panel, you know, mortality and readmission measures with, you know, where we have used similar approaches to validation without hearing some of these arguments.

But, we look forward to hearing guidance from the panels going forward, about how to proceed with optimal validation testing from your perspective.

And you know, I think in terms of discrimination, you know, we're -- we understand that clinical risk variables aren't necessarily the best risk predictors of return to hospital.

But, that does not mean that the measure doesn't have utility. Particularly as a message for understanding the implications of the more -- readmission measures that are in a payment program.

MEMBER NEEDLEMAN: Yeah, no Lisa, I totally get that. And you know, we see, you know, risk adjustment models that have R squares of 3 per -- .03, which is about what this one is, and .4, 40 percent of the variance explained.

And we just need to think about how

comfortable we are with whatever level of explained variance there is in the risk adjustment model.

And what you're telling me, is the things we know affect risk of readmission, are simply not captured in the claims data and in the billing data. And we've, you know, it's not surprising to see a low R square on this risk adjustment model.

CHAIR NERENZ: If I could just --

DR. SUTER: Yes. Much better articulated. Thank you.

CHAIR NERENZ: Just a quick time check. We've got a couple conflicting things going on.

And one is, technically we're over time. Although, the next two measures are so similar, that I think the discussion can be more streamlined.

I did just want to respond to Larry's comment. If there's -- we may have a couple of minutes for further discussion.

And if somebody has a really important point to make on that issue, I think the window would be open, because then we could do it here, and not have to do it on the next two measures.

But also, I just wanted to respond that, you know, Larry's been extremely clear and eloquent on this point. And if there's not reply discussion, it maybe that people understand the point and are sort of willing to go ahead and vote with that in mind.

So, I do think if we've got a couple more comments of things that are important, we can take a little time. But, then we've got to make sure on the next two measures we don't repeat that same discussion.

DR. MA: Thank you, David. Patrick, I see your hand up. This is -- we're way over time, so this is the last question.

MEMBER ROMANO: Well, yeah again, I just, David's point, I think that the discussion on the next two measures will be really streamlined. So, I think it's fine for us to be way over time here.

But, just to address the question of the overall performance of these models. You know, I think there is an emerging literature on using machine learning methods or deep learning methods to try to develop more predictive models for the 30 day readmissions, or if you will, the logic part of this EDAC measure.

And I think to summarize, if I could, it suggests that you might be able to get up to .7 or so from .6, by particularly including a lot of additional clinical information. Including laboratory test results, and evaluations of cardiac function, and so forth.

That's maybe a little bit of an oversimplification. But, it does -- it does suggest that there could be a significant improvement.

What I haven't seen is evidence of

whether that difference from .6 to .7 really translates into a reduction in bias in the estimation of entity level performance. And that's really the acid test that we care about.

So, if I were going to give any take on lessons, you know, for the Yale group to consider, I think that's what it would be about. Is that we know that it's possible to build better models with other risk information.

And we really want to know whether doing so would, you know, reduce bias such that these models that we're seeing today, based only on the claims data at .6, are really not sufficiently valid.

DR. MA: Thank you, Patrick. Does Yale CORE want to respond? Or, if not, we can move onto the vote.

MS. GRADY: I appreciate that last comment. We have done some work with use measures, doing machine learning techniques.

And we have seen improvement more so

on the mortality measures and the readmission measures. But, it is something that we're continuing to look into.

DR. MA: Thank you, Jackie. I think now we can move on to the vote for validity. Hannah, are you ready?

MS. INGBER: Yes. Thank you. So, voting is now open on Measure 2880 for validity. Your options are high, moderate, low, or insufficient.

And as a reminder, this is just for subgroup one. And please don't clear your response after you've voted. I'll just close the poll and then just keep it there.

DR. MA: Hannah, are you able to show your screen?

MS. INGBER: No. Not until it's closed.

DR. MA: Okay.

MEMBER ROMANO: Just to clarify, this is the overall assessment of validity?

DR. MA: Correct.

MS. INGBER: Yes. Okay. I'm showing eight results. So, I'll just close the poll.

All right. I'm ready to read the results whenever you're ready to share. Oh, thank you, Caitlin.

All right. So, we can see for validity for measure 2880, we have seven votes for high, sorry, zero votes for high, seven votes for moderate, zero votes for low, and one vote for insufficient.

Therefore, the measure passes on validity.

DR. MA: Sorry, on my screen, I don't know if it's on everyone's screen, the numbers are not lined correctly. Oh, now it is.

But we can't see -- okay. Can everyone see the full result now? Okay.

MS. INGBER: I'll read it again just for the transcript.

DR. MA: Okay.

MS. INGBER: Zero votes for high -- zero votes for high, seven votes for moderate, zero votes for low, and one vote for insufficient.

Therefore, the measure passes. Thanks, everyone.

DR. MA: Thank you. Now, we're going to move onto the next measure. Matt, you're up to introduce this measure.

MR. PICKERING: Great. Thanks. So, this is also a Yale CORE measure. It is a maintenance measure. This is 2882, Excess Days in Acute Care After Hospitalization for Pneumonia.

Just a brief description, this measure assesses days spent in acute care within 30 days of discharge from an inpatient hospitalization for pneumonia, including aspiration pneumonia or for sepsis.

Not severe sepsis, with a secondary discharge diagnosis of pneumonia coded in the claim as present on admission, and no secondary diagnosis of severe sepsis coded on pres -- as present on admission.

And similar to the previous measure, it's intended to capture quality of care transitions provided to discharged patients hospitalized for an eligible pneumonia condition by collectively measuring a set of adverse acute care outcomes that can occur post discharge within the emergency department visits, observation stays, and unplanned readmissions at any time during the 30-day post discharge.

So, it is an outcome measure. It's also using claims data at the facility level of analysis. It is risk adjusted, including 41 risk factors.

For reliability, the SMP in the subgroup, the preliminary votes for this rated it as moderate for reliability. So, this measure has passed on reliability.

But, I'll just touch on validity, in

which the measure did not pass. So, validity testing conducted was at the score level.

There was face validity. And that was assessed with a 16 member technical expert panel, in which more the 80 percent of the experts moderately or strongly agreed with the validity of the measure.

As far as the imperial validity testings, the developer conducted construct validity. It was assessed as relationships between the pneumonia measure score and the risk adjusted readmission rate group scores in the overall hospital rating scores in the pneumonia readmission measure.

The developer hypothesized a negative relationship between the pneumonia excess stay scores and the star rating readmission group score and star rating summary scores.

They also hypothesized a positive relationship between the pneumonia excess stay score and the pneumonia readmission scores.

And those hypotheses were shown with the results of the correlation analysis with negative and positive directions. And so those are also varied. And these results can also be found on page 11 of the discussion guide.

As far as risk adjustment, the developer found a C statistic of .62, and an R squared of .038.

Two social risk factors were tested and were found to be statistically significant, but did not appear to meaningly affect the hospital performance estimates, and were therefore not included in the measure.

So, I'm going to pause there and turn it over to our lead discussant, Sherrie, to mention any of the SMP concerns with this measure. Sherrie?

MEMBER KAPLAN: Is that Sherrie, Sherrie Kaplan or Sherrie Kay?

DR. MA: Yes. That's you, Sherrie. And I would just add to that, a lot of the concerns are very similar to the first measure.

So we could be brief on the similar ones.

MEMBER KAPLAN: Yes. I had myself down for 2882. But that's okay. Because the issues are pretty -- are very similar.

And they're really, the only unique thing that I would comment on, is the question of -- the question of meaningful differences.

And the developer put forward that, you know, the position that the hospitals need to look at each other's data as in, for purposes of comparison.

But, for all three measures, it looks like the idea that these measures are meaningfully different from each other, is that you can actually look at other things that would move in the direction that would establish discriminate validity.

Like it's really important. These small differences are really important, because they have, they're associated with other things.

And yes, the developers have done a reasonable job of giving us some other things they are associated with, the CMS star ratings, and the HCAHPS measures. But, it's you know, small differences are still disturbing in readmission and limitation of its own.

So, I would just like to hear a little bit more then, is there anything empirical on possibly the next round that you could give us, that has to do with either process measures, this was raised by one of the other responders as well, that could give us a little bit better sense that small differences between readmission rates except at the extremes, are helpful in discriminating hospitals one from another.

DR. MA: Lisa, are you trying to respond? We can't hear you.

DR. SUTER: I apologize. Yes, I'll -- Sherrie, thank you for that comment.

You know, I think first of all just,

you know, we do actually have a number of statistical outliers. You know, 614 out of the 46 hundred or so hospitals have fewer return days to hospital then average. And over 1,000 had statistically more return to hospital days then average.

In addition, you know, recall that this measure is -- also has a role of a balancing or a monitoring measure for the readmission measures to understand the impacts of tracking all of the return to hospital events for patients.

But, here we've heard in prior SMPs, we're hearing the response about wanting us to show comparisons to process measures. We will absolutely look into that and bring that back.

It is a shift for us, because for so long we have heard from standing committees and from, you know, research about the CORE correlations between process measures and outcomes.

And so, you know, it's a -- it's kind of a fundamental shift for us to move in that direction. But, we understand and we're hearing it.

So, you know, I will say process measures are not the priority from CMS. And you know there's been a lot of focus going forward on outcome measures.

So we have been working to do impaired validity against other outcome measures, which is sometimes challenging. Although it's getting easier since there are more outcome measures out there.

But, we hear this. And we will definitely take this back to CMS and to our group to address in the future. So, thanks so much for that feedback.

MEMBER KAPLAN: Thank you. There's just one other -- one other thought that was raised, which is the issue of responsiveness of efforts to move the needle on this.

So, if efforts to improve quality can be associated with actual measurable improvements in quality that are meaningfully responsive, that's another potential validity issue.

But, that may be beyond what's available in the CMS data set. So, thank you for your response.

DR. MA: Sherrie, I just want to ask if you can elaborate the first point a little bit for me. Because I feel like we have been repeatedly telling the developers, now we're looking for correlation analysis and the results.

We're looking -- we're really looking to see if this measure is correlated to a real valid performance measure. And that that way we will know this measure is really measuring quality.

So, how can you reconcile those two points?

MEMBER KAPLAN: Yeah. That -- it's very, very difficult. Because when you're choosing validity measures in my view, and I don't know how the rest of the committee feels, but you have to choose something that would logically be related to something that is related to.

If these EDAC measures are measuring quality, what else would people who have an excessive day stay, should -- what should those hospitals also have?

People who have core quality on one measure should also have core quality related to something else. And you get in this endogeneity group, which we just talked about, on 80, which is, well, they should also have more readmissions for other conditions.

Is that sensibly arguable? You know, or is it condition specific? I don't think there's a neat answer to that.

So, and I think that's something that the group may need to entertain, about how to choose validity measures that are both exogenous to whatever it is you're looking at, and still conceptually related.

So, I don't have a neat answer to that one.

DR. SUTER: This is Lisa, just -- Sherrie, sorry, just follow up on your question.

So, it seems like, you know, the patient's experience of, you know, physician and nurse communication and discharge information on HCAHPS domains, you know, are -- is that what you're -- is that filling that gap now?

But, -- sorry, my Apple watch just spoke to me. You know, are those comparisons what you're looking for in asking for process measures?

You know, theoretically as a clinician, that makes sense to me, right. That you should have some correlation with the patient's perception of the communication that's happening by the care team, and the information if you provided at discharge, that that would be associate with return to hospital rates.

So, that feels like we're trying, you know, we're triangulating in the direction that you want. But yet, I hear that we're not meeting your expectation.

Can you help us understand how, you know, another example that might address your concerns?

MEMBER KAPLAN: Well, I think that one could be plausibly argued. But again, the patient's experience could be one of a very small number of drivers.

So, discharge makes a lot of sense. Discharge to me, you have a conceptual relationship. And did you get the right instruction.

But excess days in the hospital, did you get readmitted because you didn't get the discharge instructions correctly?

But, there's a lot of mediating effects as -- from risk factors that, as Larry was pointing out, may have not as much to do with what happened in the hospital, as what happened outside of the hospital.

And yeah, I don't want us to go down the rabbit hole of attribution. But, things that could be plausibly argued as spending a long time in the hospital, why did you spend more then the days, you know, they expect -- that you were expected to stay in the hospital?

Puzzling through, is that just a patient experience measure? Should some of that be reflected in a patient experience measure?

Or should it be reflected in some clinical management information? I don't have a neat answer to that.

But, you know, again, the data you provided at least is one step in that direction. It's the meaningful differences that were disturbing me initially, you know.

So, can we get some sense of how much -- how much, what's the magnitude of these differences, except at the extremes, that allows us to rank hospitals one against another, and therefor do comparisons that are real meaningful?

And again, I don't have a meaningful answer. And thank you for providing the data that you did.

CHAIR NERENZ: If I just -- can I get a full agenda observation. This is rich discussion, but this is, I think, fundamentally about how could this be done differently or better in the future.

What's in front of us is how are we going to vote right now, today? And we, the comm -- the technical methods panel, really have much more to say in the future about what correlations to choose and what kind of level of correlation do you do process outcomes.

But, I think we have to bypass much of that today, just to stay focused on what are we going to vote about these measures right now.

DR. MA: Thank you, Dave, for that

anchoring. Jack, I saw your hand up and down. Do you still want to add a comment?

MEMBER NEEDLEMAN: No, no. I just apropos of David's comment. I would like to see this issue of correlational validation added to our general monthly meeting as another topic for discussion.

We don't have to do that here.

DR. MA: All right. Thank you, noted. I think that thing actually goes across many measures in this review cycle. So, we definitely needed to add it to the discussion.

So, at this point, I think I will invite Hannah to start the vote again. Again, it's the vote for Measure 2882 on validity.

MS. INGBER: Yes. Voting is now open for validity on 2882, excess days in acute care after hospitalization for pneumonia. The options are high, moderate, low, or insufficient.

All right. I'm seeing eight responses. So, I'll just close the poll. Thank you, Caitlin.

So, you can see here, the vote for Measure 2882, we have zero -- on validity, we have zero votes for high, seven votes for moderate, zero votes for low, and one vote for insufficient.

Therefore, the measure passes on validity.

DR. MA: Thank you, Hannah. And we can switch back. Thank you. Now, we move onto the third Yale CORE measure, 2881.

Again, very similar to the last two measures we just discussed. Matt Pickering?

MR. PICKERING: Thanks, Sai. So, 2881 which is Excess Days in Acute Care After Hospitalization for Acute Myocardial Infarction. Another maintenance measure in our (audio interference).

The measure is risk standardized scores of the hospital level for days spent in acute care for patients with an AMI. The measure estimates days spent in acute care within 30 days of discharge from an inpatient hospitalization for an acute care, an acute myocardial infarction, excuse me, AMI.

This measure is intended to capture the quality of care transitions provided to discharged patients hospitalized with an AMI, by collectively measuring a set of adverse acute care outcomes that occur post discharge. So, within the emergency department, observation stays and unplanned readmissions at any time during the 30 days post discharge.

This is also a Yale CORE measure. This measure uses claims as well, at the facility level. It's risk adjusted with 31 risk factors.

For this measure, as we've discussed previously with validity, a lot of the similar types of issues come up. But, I'll touch on reliability as well, just because there is consensus not reached on reliability.

And so the reliability testing

conducted was at the measure score level. The developer used a split sample reliability without replacement.

And to assess reliability comparing interclass correlation coefficients for hospitals with varying numbers of admissions. So, there's split sample reliabilities range from .23 to 0.63, with roughly three-fourths of the hospitals, those with less than 50, or 50 or less admissions having ICC values of less than .4.

The ICC was .38 for hospitals with greater than or equal to 25 admissions at .63 for hospitals with greater than or equal to 300 admissions.

And I'll just go to the validity aspect here as well. It's similar to the other measures. There was face validity conducted with a 60 member technical expert panel, with greater than 80 percent of the experts moderately or strongly agreeing with the validity of the measure.

The construct validity was conducted as well for empirical validity testing. Relationships between the AMI measure score and the risk adjusted readmission rate group scores, the overall hospital rating scores, and the AMI readmission measure.

And the developer hypothesized a negative relationship between the AMI excess stay scores and the star rating readmission group score, and the star rating summary score.

They also had signs of positive relationship between the AMI, its excess day score, and the AMI readmission scores. And those correlations can be found in the discussion guide on page nine. But, they were as predicted.

For risk adjustment, the developer found a C statistic of .6. And R squared value of .061.

And similarly to the other measures, two risk factors were tested. And they were found to be statistically significant.

It did not appear to affect the hospital performance estimates. They were therefore, not included in the measure.

I'm going to stop there. I'll turn it over to the lead discussants for this measure, which is Eric and Larry. Eric for reliability, and Larry for validity.

Maybe we can start with Eric?

MEMBER WEINHANDL: Sure. Yeah, so this is similar kinds of -- all the similar kinds of issues as we saw in the other EDAC measures and just like these.

I think that the challenge with the reliability is simply that the ICC values, generally speaking, are lower here. Which to my mind, having done lots of claims analysis, is that AMIs present fewer opportunities for accruing a large denominator then heart failure discharges and pneumonia discharges do.

So, I think that the behavior of this measure is pretty predictable. And it's not surprising that the reliability performance would be lower.

The one question that I would like to ask the developer is, having read the discussion guide over and over, and the responses which are greatly appreciated, I'm still struggling a little bit with interpretation of the figure that was included.

It's the Figure One that's on page 63 of the discussion guide, that looks at the so-called similarity between samples, and in the split sample ICC and shows the two curves sort of deviating from each other a bit as we go into higher hospital volumes.

So, I'd appreciate a little bit of the developer's perspective on that particular response. Just to help me understand my own interpretation there.

But, I'll leave it at that.

MS. GRADY: Hi, I'm going to ask that our statistician, Huihui, respond to that. And hopefully she's still on the call.

That she's actually there then.

DR. YU: I am.

MS. GRADY: Thank you.

DR. YU: Hi. We actually, I believe that lots of physicians have the -- they understand that the smaller the cohort is, the harder to sway the smaller cohort uses to equivalent samples.

So, we -- I think many people mentioned this. And which this time we just tried to present how different the two random split samples are when we -- when we -- for our very small cohort like AMI EDAC measure.

And then when we passed the similarity of the two samples, we actually used the percentage of prevalence of comorbidities. So, we compared whether or not if two different people are similar in terms of the recent comorbidities.

If they're totally different, if

there's two groups that are facing like a different level of comorbidities, we cannot say these two samples are equivalent to each other.

And then we lost the base to test the, this sample reliability, because that's the fundamental assumption, right. So, that for sample reliability, we have to test on the two, we have to test the split sample reliability on two equivalent samples.

So, this is what we did. The red dot and the red cross, and the dash line is actual similarity. We checked for correlation between -- the correlation of the comorbidity prevalence between the two random split samples.

And you can see, the red cross and the dash line, increased as the hospital level volume increased. That means when the hospital is larger, the similarities of the two samples go, and the bigger hospital will be higher.

And then, the blue circles and the blue solid line, is actually the IC -- the new

ICCs. Because some hospitals will have exactly the same volumes.

So we calculated the mean ICC code the same for the hospital with the same volumes. And then we plotted them.

And you can see as the similarity increased, the ICC increased. It's not like nothing, it's not like increased with like up to one. This actually has like a ceiling there.

So, that means that AMI still, like we are not claiming that the similarity is like too high. But the ICC could achieve one. We're saying that like because of the limit of those similarities, because of the limit of this AMI EDAC cohort.

So, if we use all the hospitals, the ICC is not impact -- is not only -- the lower ICC is not only caused by of a measure, it's also caused by the counters of this population as well.

So, if another thing about this

measure is, if we use 200, it still covered like over 80 percent of the population, of the patients. So, although with like lots of hospitals, it's not included.

However, like in terms of the percentage of patients that are covered by the measure, can be reliably measured, is up like over 80 percent. So, that's over, yeah. Thank you.

MEMBER WEINHANDL: That's a very helpful explanation. Thank you. I appreciated that.

DR. YU: Oh, thanks.

MEMBER NEEDLEMAN: This is Jack. Can I just ask for a quick clarification of how the two samples were constructed on the red, on the red zone of that chart?

I'm not seeing it in, immediately in the description that was provided.

DR. YU: So, it's the correlation. Like for example, for each sample when we split the, the random split that you have population into two random split samples, will actually have like the stratify internal by hospitals.

That means like for any hospital with at least two samples, we can randomly split them into two like samples. Right.

So, we calculate the prevalence of the comorbidity in one sample for each hospital. And calculate the prevalence of the comorbidity in another sample. And then we calculate the correlations.

MEMBER NEEDLEMAN: Okay.

DR. YU: Yeah.

DR. MA: Patrick, you're next.

MEMBER ROMANO: Sure. Yeah, so I think that unfortunately to me, this is not a close call. Because if we're going to have any kind of standardized approach to assessing measured reliability, we have to draw a line somewhere. All right.

And unfortunately, you know, this

measure as it's presented in Table Two of the submission, it falls below the line. So, with the currently used threshold of 25 admissions the split sample ICC, including the Spearman-Brown adjustment, is .384.

And that's less than for the other two EDAC measures that we're discussing. And it's less than the lowest accepted cutoff that we've discussed previously, which is around .4.

And there's nothing in the figure that really changes that. I mean, in other words it appears that the threshold, the volume threshold would have to be higher in order to get the overall split sample reliability up to .4.

So, perhaps the recommendation would be that this measure should be reevaluated with a higher threshold. Perhaps 100 is the minimum volume threshold that's needed.

But, at this minimum volume threshold, I think the measure clearly fails on reliability.

DR. MA: Jackie, other people from the

CORE, do you want to provide a response to Patrick's comment about the cutoff point?

DR. SUTER: Hey Jackie, this is Lisa. I'm going to jump in just quickly.

Thank you, Patrick, I agree this is absolutely a trade-off between the minimum volume cut-off and reliability, and the trade-off in raising the minimum case volume and gaining reliability is you're measuring fewer hospitals.

As HuiHui said, in measuring fewer hospitals, because they're small volume hospitals, you're not losing that many patients but you are losing entities being measured.

And in many ways I feel like this is a question for CMS. I think Jim Poyer is on the line and Jim, I don't know if you want to jump in about CMS's approach to deciding a volume cut-off.

If you were to go to a volume cut-off at 50 admissions, the split sample ICC is 0.402, which would meet Patrick's cut-off of 0.4. 100

admission cut-off would be 0.47, to get up to 0.5 you'd have to go above 200.56.

So, again, these are implementation decisions, not necessarily specifications to my understanding. But I'm happy to take this back to CMS if they're not on the line, we're happy to hear CMS's thoughts if they are on the line and we defer to the Committee.

Thanks.

MR. POYER: Thanks, Lisa, this is Jim Poyer, if I can speak? I don't know if anyone can hear me.

DR. MA: Yes, we can.

MR. POYER: Okay, yes, and thanks, Lisa, and thanks for the comments.

I appreciate it in terms of the recognition of the reliability concern and I think we had weighed, as Lisa had pointed out, raising the cut-off would effectively reduce the number of hospitals measured by -- if we're talking about 100 -- I believe in the hundreds.

I don't have the exact number with me, we can provide that to you. We also recognize when we had added these outcome measures and the EDAC measures were added after the readmission and mortality measures to our hospital programs.

In terms of having for the vast majority of measures a cut-off that we had to reasonably represent in terms of the vast majority of hospitals had reasonably reliable data but still in terms of that cut-off was applied across measures and measure groups.

So we weighed that in terms of rulemaking, in terms of applicability so that more hospitals would have information for the group of measures as well as a single cut-off.

And that's where we landed at 25, and I think this is in terms of, agreed, it's not at 0.4.

And obviously in terms of reliability, that said, I think for a sufficient volume of hospitals, we had made the determination in terms of we believe it would be helpful to provide this information for these hospitals where, as I think Jackie had pointed out either on this measure or other measures, it's complementary information for the hospitals and the consumers.

And recognizing both reliability as well as clarity in terms of a single cut-off, as well as in terms of providing that complementary information. And that is where we effectively had landed.

So, I hope that helps. Thanks.

DR. MA: Thank you, we need to move along. At this point, we're going to open the link to vote on reliability. After the vote we're going to discuss validity.

Hannah, are you ready?

MS. INGBER: Yes, and I apologize if there's some background noise. Voting is now open for reliability for Measure 2881, your options are high, moderate, low, or insufficient.

All right, I'm seeing the results,

give me just one minute. Okay, you can share the results, Caitlin.

As you see, the results for Measure 2881 are 0 votes for high, 3 votes for moderate, 5 votes for low, and 0 votes for insufficient. Therefore, the measure does not pass on reliability.

DR. MA: Thank you, Hannah. Next, Larry, you are going to lead the discussion on validity and since we are pretty tight on time, I would ask you to focus on the validity discussion that has not been done for other measures.

MEMBER GLANCE: So, I really don't have anything further to add.

DR. MA: Did you mute yourself?

MEMBER GLANCE: I'm sorry, I just made the comment that I really don't have anything further to add to the discussion on validity, thanks.

DR. MA: Thank you, does anyone from

Group 1 have any other comments on validity for this measure? Anyone from the whole panel have any other comments?

Okay, then I think we can go on for the vote. Hannah, are you ready for the validity vote?

MS. INGBER: Yes, thank you, so voting is now open for validity on Measure 2881. Your options are high, moderate, low, or insufficient. All right, I'm seeing eight results so I'll close the poll.

You can feel free to share. Thank you. So, as you can see, for validity on 2881 we have 0 votes for high, 7 votes for moderate, 0 votes for low, and 1 vote for insufficient.

Therefore, the measure passes on validity.

DR. MA: Thank you, Hannah. I do I want to say the results of all the voting will be shared with the SMP Members and developers at the end of the meeting.

All right, so moving looking to 3612, Matt, it's your measure again.

MR. PICKERING: Yes, so this now is on Page 12 and 13 of the discussion guide and this measure is also a core measure.

It's NQF 3612 Risk-Standardized Acute Cardiovascular-Related Hospital Admission Rates for Patients with Heart Failure under the Merit-based Incentive Payment System or MIPS.

So, this is a new measure, it's risk-standardized rate of acute cardiovascular-related hospital admissions among Medicare fee for service patients aged 65 years and older with heart failure or cardiomyopathy.

It is an outcome measure, it's using various different duties for most of its claims, Medicare fee for service administrative claims specifically.

The Roman database, it uses the U.S. Department of Agriculture Economic Research Service 2016, Area Health Resource Use File as well as the American community survey and MIPS-eligible provider files.

It's a clinician level at the group and individual analysis. It is risk adjustment inclusive of 30 risk factors, it is consensus not reached on both reliability and validity so I will just summarize reliability first.

And so the reliability testing was conducted at the performance core level and the developer performed a signal-to-noise analysis of the minimum hear failure sample patient size for the group level, the TINs, needed to achieve a minimum reliability score of 0.4 and 0.5 with a range of adequate reliability to be between 21 and 32.

So, the SMP Members agreed that the approach is appropriate but they raised several concerns which we'll talk about here a little bit.

For validity, again consensus not reached. The validity testing conducted at the performance measure score level.

There was a face validity that was also conducted and demonstrated a thorough assessment of measure face validity and value of external groups, specifically a technical expert panel, used to establish and measure development and validity.

Of the 13 Clinician Committee Members who responded to the survey, 11 of the 13 or about 85 percent strongly, moderately, or somewhat agreed that the MIPS heart failure measure can be used to distinguish good from poor quality of care.

There was also some additional concerns raised by the SMP, which we'll turn to my lead discussants to discuss on this measure.

And that for risk adjustments specifically, although there were some indicators of heart failure disparity that are not included, the model did not appear to account for the repeated measures' impact.

And there were also questions about why race and dual eligibility are not included because both could affect the outcome.

The lead discussant for this for reliability is John and for validity it's Terri and so we'll start with reliability, and I'll turn it over to John to raise some of the concerns related to reliability.

MEMBER BOTT: Yes, I'll just touch on real quickly the two areas of reliability that caught my eye.

And the first one, which resulted in my low rating, I noted that I didn't see how the measure was fully specified specifically in regard to two areas, one being the definition of E&M visits.

Yale CORE pointed out in response to my concern that indeed E&M visits were defined in the data dictionary.

It would have been nice in the IMF form if they would have said after that E&M visits to check the data dictionary because sometimes they did reference a point of the data dictionary, sometimes they didn't.

And when they didn't point to the data dictionary, I had assumed it was not further to find in there but appreciate the additional direction that it was defined in there.

The other are that I didn't see a definition for was hospice, as I think I was pointing out the exclusions. Yale CORE pointed out in their response that it's identified through the Medicare enrollment database.

However, while that supplemental information is somewhat helpful, it would have been nice to flesh that out further, what fields in the form are drawn upon, what codes, when as far as timing.

It would have been nice to have that fleshed out so the measure could be further specified.

The other issue was in regard to what

is the unit of analysis here? The measure was submitted at the individual clinician level and the group level.

When I was reviewing the test result, the discussion of TINs, I had understood TINs to be, but I was incorrect, I had understood the reference to TINs being these are group-level findings.

But in Yale CORE's further clarification, they said that basically they had rolled together individual clinical measure scores and group level scores and the results that we're seeing in the testing form is a representation of the combination of those individual level scores and group level scores.

They further went onto say that about 32 percent of those responses were of the individual clinicians.

So far in my experience with this group, I believe this is the first time where I've seen it stated that we're submitting testing results for two different units of analysis.

However, the results were given to you to ascertain how well the measure is doing and those two different units of analysis are combined and rolled up in one figure.

I didn't see that as necessarily useful for me to pull them apart to see how is the measure doing with one unit of analysis, how is the measure doing with another unit of analysis?

So, I'd be curious what NQF or the fellow SMP Members, especially in Group 1, thought of the submission of test results in that regard, where they're rolled up.

But in this case, the measure steward is asking for endorsement at two different levels but presenting us rolled-up results, essentially.

So, I'll stick to those, especially for the sake of time, those two issues I noticed and I'll open it up to whatever wants to respond at NQF and the fellow Group 1 Members.

DR. MA: Thank you, John, for your very comprehensive review. I'm curious if anyone from Subgroup 1 has other comments to add?

CHAIR NERENZ: Just a quick clarifying question. Clearly, the issue of sample size greater than or equal to 21 is crucial here I think, at least in my mind.

Are we to vote this on the understanding that this is the stipulation now in the endorsement in this measure going forward? It's only to be used where sample sizes are greater than 21?

DR. MA: Yale CORE team, do you want to respond to that question?

DR. LIPSKA: Sure, my name is Kasia Lipska. Hi, everyone, I'm happy to respond to both questions.

First, I just wanted to thank both reviewers for their comments and, John specifically, thank you for the specific feedback on where we need to be clearer in our forms.

So, just to start maybe with the TINs, under MIPS clinicians can decide whether to report as either individuals at the MPI TIN level, or as part of a group.

And the TIN, therefore, includes both solo clinicians, so those clinicians who opt not to report with other clinicians under MIPS, and also groups of clinicians who choose to report under a common TIN.

So, as John pointed out, you're right that the testing results include both individual clinicians and clinician groups, and that's consistent with how MIPs program evaluates quality.

So, as said for the TINs, with at least 21 heart failure patients, which corresponds to a minimum reliability of 0.4, about one-third or 31 percent represent solo providers, which the one MPI is defined in that one TIN.

So, it's a bit of a mix because that's

how the program evaluates quality.

In terms of the last question, what is set in terms of minimum reliability? I can't speak for the program, the program will set the minimum during rulemaking and that's during implementation.

Based on their prior use of the measures, they're certainly interested in reliable measures and typically set that minimum reliability at least at 0.4 in the MIPS program.

But the specific decisions about that cut-point for the volume is usually done through rulemaking.

DR. MA: Sherrie, you have your hand up?

SK: Back to John's point, it strikes me that, first of all, as you point out, 32 percent of the sample is solo providers. And the strategy for cutting it at 21 patients is going to overrepresent larger group practices.

But the number of people in those

practices varies, as you point out, broadly and the question then is, is it 21 patients by 1 provider in those large group practices or equally distributed across multiple providers in this group practice, et cetera?

So, who's quality are we actually studying?

And when you look at is it one provider, do all providers practice the same way across patients, the physician thumbprint, if you will, and how consistent with that is within a practice across providers is another level of concern.

So, I'm a little concerned about the over representation of group practices with a sample size of 21.

And then the question is how much variation, how reliable within a practice, how consistent within a practice or how consistent are providers within a practice? And does that need to be taken into account when we're looking at reliability?

DR. LIPSKA: May I respond or are there more questions?

DR. MA: You can respond and then Larry has another question and then we need to move on to the voting.

DR. LIPSKA: So, just quickly, I think it's a great point. I think to me this has more to do with how MIPs evaluates quality and how they consider accountability.

Is it individual accountability versus is it a group accountability? And under MIPs, when quality is assessed at the TIN level, it's either, right?

The clinician can choose to be either individual or group, and that's the unit of accountability for the quality they provide. And thus, that is what we're using in assessing the measure.

But it's going to vary, we looked at the sizes in terms of the number of clinicians, I'd be happy to provide those for you. That percentage, 31 percent, refers to already applying the cut-off of at least 21 patients.

As you noted, when minimum reliability cut-point is set higher, you're right, there's going to be slightly fewer of those solo clinicians who may still have large volume of heart failure patients.

But at a reliability of 0.5, I believe it goes down to about 25 percent. It's still a sizeable proportion of solo providers in that sample.

DR. MA: Thank you, Larry, you are up. You're on mute.

MEMBER GLANCE: Sorry. That was a great discussion.

I just wanted to point out that when we set the threshold, whether it be 21 patients, 50 patients, 100 patients, we're still calculating the median, in this case, signal and noise ratio for the entire group, which includes physicians or clinicians with 100, 200, 300, possibly more.

So, if you were to limit the signal to noise ratio calculation for, say, the low-volume providers, say arbitrarily between 21 and 50, it's very likely that your signal and noise ratio would be much, much lower.

So, the problem that we're having in terms of trying to pick a threshold for reliability is in terms of how we report it.

So, when we choose a different lower limit for the threshold, we're always going to have a higher and higher signal to noise ratio.

But it doesn't change the fact that no matter where we pick the thresholds, in all likelihood the low-volume providers, even if we pick it at 20, 30, 40, 50, in that group of low-volume providers we would still have, quote, unquote, unacceptable reliability if we use a threshold of 0.4.

And I think that part of the reason

we're struggling with thresholds is because we've all decided that we want not just a single number but we want know some additional information.

And whether that be the inter-quartile range or whatever it happens to be, it's going to give us different information.

So, I think the problem also that we had with the commission measure is that in reporting the results of the split sample reliability testing on a low-volume group that of course their measure of reliability was going to be unacceptably low.

And I think here, again, I just want to reiterate the fact that if you move the threshold down, you may have an acceptable reliability threshold that you've achieved but it doesn't really reflect the reliability of the low-volume providers.

DR. MA: Do you want to respond?

DR. LIPSKA: I just want to clarify, because I hope that I'm following all of your excellent comments, that minimum threshold means that's the minimum, right?

So, for 21 heart failure patients the minimum reliability is 0.4 for those smaller-volume providers but the median is 0.6 so we have at least 0.4 in that sample.

MEMBER GLANCE: What was the upper limit for the number of heart failure patients? The low limit was 21 and what was the upper limit in terms of volume?

DR. LIPSKA: We set the minimum volume but not the upper limit. But we've looked at a variety of thresholds.

MEMBER GLANCE: So, I guess my point is when you set the minimum threshold at 21 heart failure patients, it goes all the way to the maximum. So, the signal to noise ratio reflects the high volume, intermediate volume, and low volume.

And to really know what's going on in terms of reliability for the low volume, you would need to set the intervals between 20 and 50 or 20 and 75.

CHAIR NERENZ: Although I think I was hearing earlier that at 21 exactly the reliability is 0.4, if I'm understanding this correctly, and then it goes up from there to create the median.

And obviously the median is reflective of the whole range of possible sample sizes. But what I thought I understood, the reason 21 was chosen is that's where you hit 0.4.

DR. LIPSKA: That's exactly right.

DR. MA: Thanks for that clarification. At this point, we're going to move on to the vote for reliability for this measure.

MS. INGBER: Thanks, voting is now open on reliability for Measure 3612. Your options are high, moderate, low, insufficient. Okay, we have all eight results.

Caitlin, feel free to share the

results. As you can see for reliability for Measure 3612, we have 0 votes for high, 12 votes for moderate, 3 low votes for low, and 0 votes for insufficient, therefore, the measure passes on reliability.

DR. MA: Thank you, Hannah. Now we're going to move on to the validity discussion. Terri?

MEMBER WARHOLAK: Good morning, well, morning here.

So, first of all, I'd like to thank the Yale COREs for relying on to our comments and questions. You did a great job of addressing each one of them individually.

I think that I'm going to ask a question for myself and then I'll give a couple more items that I think I'd like to see touched upon with the larger group.

But for myself, it looked like to me, as is NQF policy, that a measure can be approved for validity the first time with only having face validity. I'm not sure --

DR. MA: That's correct.

MEMBER WARHOLAK: I'm not sure I'm such a fan of that but that's what it is. However, that just brings up some additional questions.

So, there was the technical expert panel as well as the clinician panel and one of the things that I'm thinking about is that there were for the technical expert panel only 12 of the 17 active members replied.

So, it looks like we have some response bias in the face validity. Also, too, the next thing I wanted to get some response to, just to satisfy my own curiosity, it is mentioned that the technical expert panel suggestions are iterative.

And so after the reviewers looked at this particular set of data, what edits were made to the measure? Or what kinds of things did the technical expert panel ask for?

I know they have some of their responses here but it doesn't tell me exactly what they wanted to be changed.

And then the same thing with the clinician panel because I was kind of struck by the fact that even on the panel not everybody agreed.

So, if you guys could speak to what kinds of things did they want you to address that perhaps you did put into your risk adjustment model in a measure or inclusion or exclusion criteria. So, I think that would be interesting to note.

Also, too, I think that I'd still like some additional comments on the issue of not including race in the risk adjustment panel or in the model.

And then also some comments on the really low R-squared. And I think otherwise, the responses really help so thank you.

DR. MA: Does the developer want to

provide a response now?

DR. LIPSKA: I'm happy to. Thank you for your questions.

So, in terms of the potential response bias from the TEP, as I think we mentioned in the response, this was a multiyear, multimeasure TEP and not all Members were always that active.

But I do not believe all of them were active. We can't tell you for sure, I would try to look at the notes to find out how many had dropped out by then. So we reported how many were really on the TEP and how many responded.

That being said, again, it was a multiyear, multimeasure TEP that helped us on this measure.

With respect to the feedback, yes, so this measure underwent multiple revisions and feedback both from the TEP and from the Clinician Committee, which was composed of both clinicians as well as heart failure society Members helped us bring this measure along.

And there were multiple changes made that I'm not going to be list all of them but things in particular that clinicians and TEP Members were concerned about are some other things to what you all have pointed out, which is there are going to be heart failure patients who have advanced heart failure who may be clustered among specific clinicians and it may be difficult to account for the risk.

And, therefore, we were very careful trying to tease out those kinds of patients, so those who were at the high end of risk. Some of them, many of them, were excluded for those reasons we could not account for that risk.

So, patients would transplant those with implantable devices who are very high-risk for admissions. Those who have end-stage renal disease was also a change that the TEP had asked for, that patients with end-stage renal disease be removed from the measure, that systolic heart failure be taken into account, which also portends a poor prognosis.

So, we're very careful to tease out all the risk factors that may increase the risk of hospital admission, and also those patients with those risk factors may be clustered among advanced heart failure providers.

There was something more but I don't remember, I think I touched on most of it.

DR. MA: Thank you, anyone else from Group 1? From the whole panel, no additional comments at this point?

DR. LIPSKA: I think you were asking about race, I didn't answer that. But we do not adjust, per CMS policy, none of the measures are adjusted for race.

I think we provided a response to that as well.

DR. MA: Thank you. Patrick, you had your hand raised?

MEMBER ROMANO: Yes, just to briefly say that the fact that five Members of the TEP went missing is sort of suspicious when we're relying on entirely on face validity.

And I guess we can accept your explanation that they may have drifted off and lost interest or whatever, but it is a little bit of a yellow flag when we're basically asked to evaluate only face validity.

DR. MA: Thank you, Patrick. Do you want to offer additional response to that comment, Kasia?

DR. LIPSKA: I wish that I could bring them back and make sure. We had both the TEP and the Clinician Committee that contributed to the measure.

It was effectively a three-year TEP so I'm not totally surprised that some of the Members were missing by the end. We did also survey the Clinician Committee and provided those responses.

I think that should be reassuring but I can't tell you, I don't have the exact numbers of where people went.

But I think that given it was this three-year, multimeasure TEP, I hope that provides you with some reassurance.

DR. MA: Thank you. Now it's time to vote on validity.

MS. INGBER: Okay, great, thank you. So, the voting is now open on Measure 3612. As a reminder, your options are moderate, low, and insufficient since only face validity was submitted.

I'm seeing eight responses. All right, feel free to share this screen, Caitlin, thank you.

As you can see, it's a little small, for validity on Measure 3612 we have 6 votes for moderate, 2 votes for low, and 0 votes for insufficient. Therefore, the measure passes on validity.

DR. MA: Thank you, Hannah, I think wraps up the first batch of the discussion.

We're going to take a ten-minute break and we'll resume at 2:30 p.m.

Thank you, everyone.

(Whereupon, the above-entitled matter went off the record at 2:21 p.m. and resumed at 2:35 p.m.)

All right, we are at 2:35 p.m., and Patrick, thank you for letting us know. We should be more mindful the next time to put together the agenda.

I do want to say we have way too many measures for the discussion this time. Okay, welcome back, everybody, we can move on to the next slide. Matt, you are up.

MR. PICKERING: So, now we have our next measure, Readmissions Measure 3188, 30-Day Unplanned Readmissions for Cancer Patients.

And so you can find this starting on Page 11 of the discussion guide and going onto Page 12. The developer of this measure is the Alliance of Dedicated Cancer Centers and this is a maintenance measure.

It is 30-day unplanned readmissions for cancer patients, the measure is a cancer-specific measure.

It provides the rate at which all adult cancer patients covered as fee for service Medicare beneficiaries have an unplanned readmission within 30 days of discharge from an acute care hospital.

The unplanned readmission is defined as a subsequent inpatient admission to a short-term acute care hospital, which occurs within 30 days of the discharge date of an eligible index admission and has type of emergency or urgent.

So, it is an outcome measure, it uses claims as its data source, the level of analysis is at the facility level. It is risk-adjusted with 11 risk factors and for reliability the SMP subgroup did pass the measure with a moderate rating for reliability.

But for today we'll be discussing

validity. As you can see, it was not passed and so before I turn it over to my colleague, Patrick, to talk about some of the concerns, I'll just touch a little bit on the testing.

So, validity testing was conducted at both the critical data element and measure support levels.

The developer created a cross-block of ICD9 to ICD10 for codes used to define the denominator, the denominator exclusions, numerator exclusions, and certain risk adjustment variables.

The correlation between 30-day unplanned readmission for cancer patients and CMS's hospital-wide all-cause readmission measure, or NQF 1789, was 0.255.

It was found to be statistically significant for 2412 hospitals for which the data on both measures were available.

So, I'll just mention briefly before I turn to Patrick, several ICD Members did raise concerns about the endogeneity nature of the correlation analysis and that is that the denominator for the hospital-wide all-cause readmission measure includes patients with cancer, and the same readmissions are in the numerators for both measures.

So, some of the similar types of concerns we've heard on some of the other measures here as well.

So, I'll stop there and, Patrick, I'll turn it over to you if you wanted touch on any other concerns that were noted by the other SMP Members.

MEMBER ROMANO: Sure, so I'll focus on two issues.

The first is the issue about the entity-level validation and as Matt's described, they justify entity-level validation based on the correlation between this measure, which is a cancer-specific measure, and the hospital-wide readmission measure.

And of course, the reviewers pointed out the inherent endogeneity in that correlation, and the developers response was basically two-fold.

They said that the hospital-wide readmission measure, X-group patients admitted for medical treatment of cancer and, of course, it also excludes cancer centers, the PPS exam cancer centers.

So, I don't think the latter issue is relevant here because, in fact, the data they're presenting for our assessment of validity is based on all the hospitals that are in the PPS.

So, that's what we have to assess before us.

Now, the question about hospital-wide readmission measure, it excludes patients who are admitted for medical treatment of cancer. But I would submit that most of those patients are elective admissions anyway.

They're patients coming in for

scheduled chemotherapy regimens or ablation regimens, and therefore, they're probably omitted from this measure as well.

The point is that even if there's some zone of non-overlap, there's so much overlap in the denominators between the measures that it's hard to get too excited about a correlation of the level that we see here.

So, this is perhaps the correlation level that we would expect just randomly from taking out a subset of the data.

Also, to be specific about one other thing because I did go back and look up the hospital-wide readmission measure in their specification, they include patients admitted for surgical treatment of cancer but they also admit patients with cancer who are admitted with other diagnoses.

And that's another area of overlap between these two measures. So, we have this problem of the choice of the measure for construct validation.

Again, a more persuasive response would have been the response that Yale provided earlier today, where they provided clear correlations that remove those inherently overlapping measures.

The second point I'll just mention is about the risk adjustment model and there are a couple things that were a little squirrely about the risk adjustment model here.

At least they raised the yellow flag, which is typically we look for risk adjustment models to be limited to factors that were present at the start of care.

And the model in this case includes a length of stay in the indexed hospital, it includes the use of ICU services in the indexed hospital, and it also includes prior hospitalizations.

And it also includes discharge to hospice care. So, it's just a violation of the standard of precept to be including things that are a part of the process of care in the hospital.

In particular, length of stay, of course, is a proxy for severity of illness but there's also a lot of data out there about how there might be correlations between the length of stay in the indexed hospital and the likelihood of readmission.

So, those are maybe the two key factors, the construct validity and the selection of variables for the model, particularly variables that are based on data after admission.

DR. MA: Thank you, Patrick. At this time, I want to invited the developer to provide a response.

MS. McNIFF: Yes, hi, this is Kristen McNiff, are you able to hear me?

DR. MA: Yes.

MS. McNIFF: And I have others representing Alliance of Dedicated Cancer Centers who all asked to join into this measure as well.

So, the first I think is related to use of CMS's hospital-wide all-cause readmission measure for the measure level validity testing. And I certainly appreciate the concern.

I do just want to recognize that the specifications for that measure will exclude any patient admitted with a principal diagnosis of cancer or, as mentioned, those who have a diagnosis of cancer and are undergoing a surgical procedure.

So, that would not certainly represent all volume of cancer patients being treated in the hospital.

And as the specifications note themselves for the hospital-wide all-cause readmission, this is a distinct population that, again, is expected to have a different course of care, a different readmission profile and that is why they were excluded and that is why we have addressed them separately in a unique, independent quality measure.

So, while we don't have access to the data for the PPS hospitals to be able to fully assess where there's overlap and where there's not, as other developers may have access to, we do believe that is the best measure we have for correlation.

Unfortunately, in oncology specifically there's just a dearth of outcome measures and not a whole lot of related process measures.

So, it was determined for the original submission that this still was the most persuasive measure to look at for validity testing, and by others from the ADCC team to reflect their comments, especially.

I'm not sure if Dr. Fields would like to add there?

DR. FIELDS: I would just reiterate what Kristen said. I'm the Chair of the Cancer Measure Committee and unfortunately, there's a dearth of outcomes measures and a dearth of cancer measures.

And this seems like important data that would be actionable for treating our patients in order to improve quality and outcomes. So, I'd just reiterate what Kristen said.

DR. MA: Thank you, Karen. Does anyone else have a comment, either from the group or from the panel?

MS. McNIFF: I can address the second issue as well, if you like. For this maintenance submission we used the same risk adjustment variables that were used for adjustment for the original submission of this measure.

We specifically had a process to have a clinical group of experts on a variety of Committees to actually take a look at those.

We reviewed the literature, however, because the risk adjustment model was performed well and was accepted for the first submission, we wanted to minimize actually the amount of change that we made to it.

So, therefore, we did not address the variables, again, except to make sure they were still consistent with the literature and were reflecting what we know to be associated with readmissions from the literature.

As you saw from the submission, we did conduct the fully analysis around the model, again, with the updated data and reported how ultimately the model was changed for main submission.

But again, we recognize the fact that this model had previously passed, was previously considered acceptable, and thus, wanted to actually minimize the changes that we made for maintenance.

I don't know if others on the team want to comment on that but that was certainly our approach.

MEMBER ROMANO: I'll just add that I think you report a c-statistic of 0.71 and as we discussed previously, this is a unusually high statistic for readmission models.

And I think part of the reason it's so high is because you're adjusting for things we don't usually adjust for in readmission models, such as the length of stay and the indexed hospitalization of whether the patient was discharged to a particular setting of care, and so forth.

The concern is, for example, that sending patients out of the hospital too quickly might be one of the factors that leads to higher readmission rates.

And of course, it's very hard to parse that out from the fact that longer length of stay is a proxy for greater severity of illness. But again, the Yale CORE Group has always kept those kinds of post-admission factors out of their response.

MS. McNIFF: I invite others on the team to weigh in, however, I'm not sure if there's additional comment at this point.

DR. MA: Do other SMP Members have any other comments either for the developer or for Patrick? Larry?

MEMBER GLANCE: I agree with the point that Patrick is making, that they should not be including length of stay as a risk factor in the model.

That really is a significant threat to the validity of the risk adjustment.

DR. MA: Zhenqiu?

MEMBER LIN: Sorry, I just want to say I agree with Larry and Patrick because you want to set it off at a point when patients come into contact with the provider.

What happened, how you need to care for the patient, right, we try not to account for what they did, and that's part of what we're trying to evaluate.

DR. MA: Thank you, Zhenqiu. If there's no other comment at this point about validity, I would ask Hannah, please do a roll call for Subgroup 1.

As one Member had to step out this afternoon, we any make sure if we need a quorum at this point.

MS. INGBER: Yes, happy to. Eric Weinhandl?

MEMBER WEINHANDL: Present.

MS. INGBER: Sean O'Brien?

MEMBER O'BRIEN: I'm here.

MS. INGBER: I believe Sherrie Kaplan left? Yes? Okay. John Bott.

MEMBER BOTT: Yeah, I'm here.

MS. INGBER: Larry Glance?

MEMBER GLANCE: Here.

MS. INGBER: Terri Warholak?

MEMBER WARHOLAK: Here.

MS. INGBER: We heard from Patrick Romano. And Dave Nerenz?

CHAIR NERENZ: Here.

MS. INGBER: All right, we have a

quorum, thank you, everybody.

DR. MA: Thank you, Hannah, now we can pull up the vote link for validity.

MS. INGBER: Yes, thank you, voting is now open for Measure 3188 on validity. Your options are high, moderate, low, insufficient. Just waiting for one more.

Okay, thank you, everyone. So, regarding validity for Measure 3188 we have 0 votes for high, 2 votes for moderate, 3 votes for low, and 2 votes for insufficient.

Therefore, the measure does not pass on validity.

MEMBER ROMANO: Sai, could I ask a question?

DR. MA: Yes, please.

MEMBER ROMANO: Could you clarify?

I remember there was some discussion at the beginning just to clarify in the case where at endorsement maintenance and the measure was previously endorsed based on data element validity.

And the measure developers bring back that evidence of data element validity but we decide that the entity-level validity doesn't meet criteria.

Where does that leave us at maintenance?

DR. MA: If I understand your question right, you would have just followed the same algorithm on our guidance to evaluate the validity at the performance score level.

So you would start from high and go down through other validity reveal criteria like risk adjustment on measure specifications and other subcriteria. Is that what you're asking, Patrick?

MEMBER ROMANO: Yes, I was just asking because we weren't voting on the subcriteria so it makes it a little bit confusing.

DR. MA: Right, it's overall voting on the validity after you consider all the test and methodology and other factors. And the threats to validity exclusions, et cetera.

Okay, what are four minutes ahead of time for 3622. Before we start, I just want to make sure the developer and the steward from the Human Services Research Institute is at the meeting at this point?

DR. LI: Yes, we are here.

DR. MA: Okay, great, thank you. All right, then we can move on to 3622. At this point, I would like to invited my colleague, Sam Stolpe, to lead the discussion.

DR. STOLPE: Very good, thank you, Sai, and hello to all of my friends and colleagues on the Scientific Methods Panel.

I think this is one of my favorite times of year when I get to have the opportunity to spend some time with you and talk through some of these issues.

I always learn a lot from these discussions so thank you all for your time and contributions.

A little bit about 3622, just a couple of background items before we get too far into just running through the slide and before I hand it over to Dr. Nerenz. I wanted to point out just a couple of process items.

So, first, as a reminder, this is an instrument-based measure and as such, it has more strict requirements than many of the other measures that we have at NQF.

So, that includes the testing requirements at both the data element and score level for both reliability and validity in order for the measure to pass.

So, of course, this measure passed reliability so I won't spend too much time there but when we get into the details around validity, I want you to especially pay attention to.

The other thing that I want to point out and it's something that we haven't talked about a lot since we had all the CAPS measures come through in spring of 2019. So, it may be helpful for the Committee to be reminded.

For many instrument-based measures, we have multiple rates that are reported and what that means is we actually need to consider the data element validity and the score-level validity, et cetera, all the testing for each individual measure rate component.

So, potentially, we could, for this measure that has 14 different measures tucked underneath the auspices of 3622 title, be voting on any individual measure within it.

How this would look like on the way that we represent 3622 would be that if there is one component that we feel doesn't meet NQF requirements associated with validity and reliability, that would not be listed amongst the elements that we have endorsed.

Before I go any further about that, I just want to let you know what the process would be. And it's that if you wish to pull up any individual component of it for separate voting, we can do it that way.

So, if you feel like you would vote down validity because of one specific element, then that's something you should discuss as a panel so that you can go ahead and vote appropriately.

Any questions about that before I go any further? I want to make sure we resolve those.

DR. MA: Did you mention there are actually 17 measures?

DR. STOLPE: my understanding is that there's 14.

DR. MA: 14, sorry. So, we could either vote as 1 or we could vote on each one of the 14 measures separately. The link is right there but the utilization of the SMP Members' decision.

DR. STOLPE: Okay, great, well, if at any time during our process you have any questions about that, I'm here to remind you about how to conduct the process and I'm happy to help.

A little bit of background about 3622 before we get too far down the road, this is the National Core Indicators for Intellectual Disabilities and Developmental Disabilities inside home and community-based service settings.

As you may know, in the United States Medicaid home and community-based service waivers are the largest providers of long-term services and supports for people with DID.

And what that means is there's over 2.5 million individuals receiving HCBS doing optional 1915(c) or Section 1115 waiver and nearly 1.2 million that received optional personal care state planned services.

And 600,000 individuals receiving home health state planned services, which is the sole required benefit.

There's fewer individuals receiving

HCBS through the relatively newer state plan options including Section 1915(i) in Community First Choice.

But nonetheless, we're talking about a very large population of a traditionally marginalized and disenfranchised group of Americans.

Joint federal and state Medicaid HCDS spending totaled $92 billion since this fiscal year 2018 with nearly all spending under optional services.

The same thing to keep in mind, especially relative to this measure, is that per-enrollee costs are highly variable, with individuals with intellectual and developmental disabilities constituting the largest spend.

The average cost is $46,000 per annum. Per annual enrollee spending, for example, is quite a bit lower for seniors and adults with physical disabilities, whereas it's right around $16,000 per annum.

CMS allows dates a wide flexibility in designing ACCDS waivers so this is resulted in a fairly large disparity across states and services.

In 2018 NQF convened a Committee for Person-Centered Planning and Practice Committee that included a wide range of stakeholders with experience in LTSS and person-centered planning.

And specifically in recognition of the variability of quality across states the Committee called for standardization and for utilization of capital-looking measures as both the state, the agency, and the provider levels within LTSS broadly and ACCDS specifically.

Now, the activities of this Committee included an environmental scan of LTSS measures specifically for persons that are planning, among which the national core indicators emerged and were discussed at length by this Committee.

So, this is an echo of a call to action from that particular Committee to see more measures of accountability specific to this area brought forward on the report. NQF is delighted to see these measures brought forward for consideration of endorsement.

The measure description goes through and describes the national core indicators for intellectual and developmental disabilities in HCBS as part of a survey.

What we're looking at here is essentially an annual cross-sectional surveys of adult recipients of state, developmental disability systems supports and services.

This has been around for nearly 25 years and has quite substantial reflections in the literature of testing over time and it's very well established. Currently, there's 46 states MDC that are participants inside the NCI program.

So, inside of this measure, the survey instrument itself, there are 4 key domains, 14 measures in total, 5 measures inside of the HCBS domain, which include areas such as community, job, goal, person-centeredness, meaning proportion of people who report that their service plan includes things that are important to them.

A portion of people who express they want to increase independence and functional skills around their activities of daily living, et cetera.

There's another demand around community, inclusion, yet another around choice and control. And one domain that contains a single measure around the proportion of people who reported that their personal space is respected in the home.

So, a couple of other things to point out. This measure is a pro-PM and that is risk-adjusted and the developer presented quite a bit of both reliability and validity testing around the measure itself.

So, the ratings for reliability were three for high, three for moderate, two for low, and one for insufficient. So it did pass with a moderate rating.

For validity, however, there were two votes for moderate, three for low, and four for insufficient. The developer in what they presented for their submission was testing at two levels.

First at the data element level, the developer suggested that interviews where they were asked to give formal feedback, interviews to ensure the individual interview validity was testing at the data element level.

They also provided a list of seven references for studies investigating the relationship among NCI data elements and testing hypotheses about expected associations.

Now, traditionally, those would be summarized within the submission and the Q results associated with each of those studies would not be presented as a list of citations.

So, some of the SMP raised that as a

concern.

At the score level, the developer reported Pearson Product Moment Correlation coefficients among the 37 states' performance scores for the 14 IPS items with scores ranging from 0.345 to 0.763, which suggested a moderate to high correlation between the individual items.

However, there were some concerns raised for external measures of validity being used to provide somewhat of a stronger connection to a comparable quality concept.

With that being said, I'll hand it over to Dr. Nerenz as the lead discussant here to summarize any further concerns. Dr. Nerenz? Oh, you're on mute, sir.

CHAIR NERENZ: Not for long. There we go.

Oh, thanks. The -- that is, we've had all day and, thanks, to the staff, great summary. I won't take too much time and add much.

As we noted, earlier in the day, this

one is really, really a hard one to do, because it's, essentially, 14 separate measures. And I know, in doing my preliminary evaluation, I was doing some color coding and tables in the submission to say, okay, here are the green one, look okay, here are the yellow ones, here are the red ones.

But, in the end, we're asked simply to provide one, one vote, and so that's what we did and, and I'm still not sure, here, this afternoon, whether it's going to be feasible to get into any individual voting, because we're going to have to, kind of, unpack the whole thing, basically, in real time, to do that, so we'll see what my colleagues want to do with this.

The only couple of things I'd emphasize, in addition, to what you said, you know, these are state-level measures and, presumably, they're measuring the quality of state-level programs.

And, particularly, in the issue of

validity, what I was struggling to find, in the materials we were given, and then, I, I appreciate the additional information we got, now, in the response, is, you know, sort of, what's the theory of quality here?

That, what are these programs supposed to be doing, why do these measures reflect something important that we can define about their quality?

How do these measures not reflect, sort of, the confounding influence of other things? And then, I was trying to filter that all together and make judgments about it.

So for example, in the response -- and, and I'm now going to, sort of, tee this up so that -- I have to change screens here. So developer can respond to this.

On page -- I'm sorry, the pages don't show up in my PDF. We have, sort of, a table of responses about, you know, the validity and, and the first item in the, in the table is in reference to the community job goal, which is the first measure of the 14.

And, sort of, to establish validity, the argument is that, because urban settings provide greater job opportunities, you'd expect a correlation, at the state level, between the percent of people, who live in urban areas, and the score of the measure and, indeed, there is such a correlation.

My question is, does that establish validity of the quality of the program, or is it, actually, a statement about a confounder that ought to be adjusted for, because urbanicity isn't really, to me, a dimension of quality.

So when I see this, I'm actually -- I'm, frankly, not more convinced of the validity. If anything, I'm saying, here's something that, perhaps, belongs, if it isn't, already, in an adjustment model.

And I could go on, but I, I think we've got this question of -- and, and this is a classic question in, in multi-item surveys of this type.

When each individual items or, or subsets of items are brought forward, as measures, and then the developer seeks to establish validity, by looking at the inner correlations among them.

I'm always nervous about that, because you've, you've got some overall positive, negative response biases, you've got some situational context that might affect things, for example, this urbanicity thing I just mentioned, and it, it's just tough.

I know, it's hard for the developer, it's hard for us and, perhaps, as we turn to the developer, here, the focus can be on, why exactly does the pattern of relationships that you have shown us establish the validity of each of these measures?

And, I understand, it's, sort of, bootstrapping, but -- and, and use urbanicity, if you want to, why does that correlation establish validity, as opposed to the presence of some kind of confounding factor?

So -- and, by the way, I'm the one, who said a bunch of these are outcomes. I understand what the NQF guidance says about PRO-PMs.

I'm sorry, some of these are still not out of context, but actually that works in the developer's favor, because you don't have the -- you don't have the same obligation to risk adjust some process measures, as you do outcomes. I would expect more in the area of risk adjustment, if I thought these were truly outcome measures.

DR. STOPLE: All right. So, Dave, my question for you is, you, is you seem to indicate that you wanted the developer to respond to that, is that where you'd like to take the conversation?

CHAIR NERENZ: Well, I, I'm sorry that this is confusing, it's just that the, the whole bundle we have is tough. I think the, the, the main voting issue we have, in front of us, which we have to keep front and center in our discussion, is the mixed ratings that we got on validity.

And, I think, somehow, before we close out this afternoon, we're going to have to, either, re-vote that, or somehow, we're going to have to figure out how to re-vote individual measures, within this set, so I should've stated that more clearly.

I think the focus here should be on the evidence of validity. It should be on the fact that, our Panel, in its preliminary ratings, was kind of all over the place, in our assessment, and I'd like the developer to guide us to, how do we come to a better, hopefully, a more coherent conclusion about this?

DR. STOPLE: Yes, Dave, I think you've got some really great questions in there that, I'll turn over to Dr. Li, in a moment, to, to jump in and address.

But, I would, I would propose a slightly different order in, in us getting to the, the question around validity. First, there were a lot of votes around insufficient, and I, the reason that I'm concerned that that showed up is that, perhaps, what the developer submitted just doesn't check the box for all of the measures that were submitted.

Meaning that, we need to demonstrate, both, data element level and score level validity and there seem to be some questions in the mind of many Members of the Panel, whether or not that was the case.

So if that's not true, then that's the first series of advices that we need to give to the developer is how to, how to make sure that they're checking the boxes, to get a complete submission.

If we are satisfied with the, at the sufficiency, then I think we get to the next level of questions, Dave, and that gets to your point, which is, you know, are, are we actually measuring what we think we're measuring?

CHAIR NERENZ: Yes. And, and thanks for that and I'll just have to check with my colleagues here, on the fly, to be sure.

I, personally, did not have concerns about data element validity. That did not affect my rating. I have much greater concerns about the entity-level validity, which means, the state program-level validity.

DR. STOPLE: Okay. So -- so perhaps we can put a brief peg in the, to the question that you're proffering, here, Dave, and move to any of your other colleagues' concerns around the validity of the measure.

I, especially, want to hear, from Members of the Panel that, have voted insufficient.   Do we feel like we met the requirements associated with a full submission?

(Pause.)

DR. STOPLE: Patrick --

(Simultaneous speaking.)

DR. STOPLE: Patrick, it's --

MEMBER ROMANO: Yes.

DR. STOPLE: Okay.

MEMBER ROMANO: Yes.   I mean, I, again, I, I think that, the, the problem that we're all struggling with and, and, maybe, the developers can guide us through this.

But, the problem we're all struggling with is that, we, we're, we're trying to understand that there are 14 measures that are being presented here and, the 14 measures and, and, in Exhibit 3, the developers present various correlations.

But I, I think what we're missing is a clear explanation of what the quality construct is and, why these specific correlations have been proffered, as providing evidence for validity. It may be that, some of them are fine. But, we're, we're missing the framework.

DR. STOPLE: Perhaps that's a good place to start then, in our discussion with the developer. Did you want to jump in, Dr. Li?

DR. LI: Sure. First of all, thank you, SMNP, and thank you, Dr. Stople, and Dr. Nerenz, Dr. Romano, for your comments.

We have come into this, with a perspective of trying to take our existing measures and fit into the NQF perspective and approach and, it has, as you can observe that it's, it has not been a very smooth fit, so far.

We recognize that, there has been a misfit of approach, here. Initially, we understood the, the, the validity requirement in the passing attachment to be, present some kind of a correlation to be established internally that these measures will hold up with each other and that's what we did.

And, later, after receiving the feedback, from the Committee, we realized that that, by itself, is insufficient to stand on its own.

So in the response we have provided different, a different perspective, I will say. Because, even though we still presented some correlations, now, we're emphasizing much more the theoretical perspective of it and that's what Dr. Romano had just alluded to that, we're missing a theoretic, theoretical connection there.

So I would point everybody's attention to Issue Number 2, the response and the table and, and just say that, we are attempting, here, to provide a direction of, of our correlation analysis, to show that, we're not just randomly creating a matrix of things seemingly associated with each other, but also, as we talked about, previously, in other measures, that there, there has to be a logical reason for those to be correlated and that's what we, kind of, tried to present in this table.

So that, that's a, something to

consider for the Committee. I hope to reassure on, why we're using correlation, in this way. And, I don't know if other teammates might want to jump in here, to explain more about this, these correlations?

But, we can, also, move to answer a little bit of, about Dr. Nerenz's question, earlier question, about whether external -- there, there might be factors that might be external to the state, or external to the, the state programs that could be confounding.

CHAIR NERENZ: Well --

DR. LI: So I, I guess, I'm not sure the direction that the Committee wanted the conversation to go, whether to move to that, or stay on this topic a little bit longer and deliberate?

CHAIR NERENZ: Well, if I could just ask you, to go a little bit further?

DR. LI: Yes.

CHAIR NERENZ: Because, you know, to me, the table that we have, here on the screen, it addresses the heart of my own concerns, about this particular measure, about validity. And you can pick almost any row, here, you want.

DR. LI: Okay.

CHAIR NERENZ: In all these cases, you're explaining, why you think, two measures should be correlated with each other. In, almost, no other cases -- in none of the cases, do I see a clear statement, why this relationship reflects quality.

I understand, why Measure 10 might be related to Measure 3, I understand that. But, why does that relationship say something about quality? That's what I'm missing.

DR. LI: Got it. That's a -- that's a great point. And, I would point to the HCBS Report that NQF has produced and, back in 2016, September, about the conceptual framework.

So in that, in that framework,

the -- there, there is a clear definition of high-quality HCBS, so some of the -- there are a list of, of criteria that, that we considered.

And, again, this is -- we're -- what we're attempting to do, here, is to take existing measures that have already been developed, prior to our NQF-endorsement work, and try to fit in, fit in this perspective.

And, some of the examples of a high-quality HCBS might be to, first, provide a person-driven system to optimize, as individual choice, and that can be seen, here, in our measures that are in the choice and control domain.

So essentially, we have two risks over-simplifying things, a bit, but, here, the higher score you get, with choice, the, the theory is that, the, the higher quality the system is providing to this, to this individual.

And the same goes for the other criteria that the, that the report, the NQF

Report on HCBS has listed, there is a focus on social connectedness, an inclusion of people, who uses, who use HCBS.

And, again, it goes back to our, our submitted measures on person -- Social Connectedness number CI-1, the proportion of people, who reported that, do not feel lonely, often, that's a, that's a question that directly addresses that, that focus and that criterion of high-quality HCBS.

So what we're proposing here is that, with -- even though, we don't have a clear clinical model, per se, to indicate the cause and effect of, between, between the different characteristics and the outcomes that we are measuring, we are showing that, on a, in a framework level, if you look at the list of things that are in this high-quality definition, the definition of high-quality HCBS, you can see a clear alignment between the measures that we're submitting and the, and those, those criteria.

And I just want to say that, it -- thanks, for letting us know that, this, this could have been clearer, in our form. And, I realize that, within this format, we didn't really have a chance to, to open up and explain things, thoroughly, so I thank, everybody, for your patience, with reading our submission.

And, thank you, Dr. Stople, for, in the beginning, for introducing us and contextualizing the discussion and providing a platform of understanding, so we can, so we can base those discussions upon.

And, in, in the submission, if we move forward, from here, we'll definitely include a more thorough discussion on, why this matter is the value of things?

And, why we think our measures align directly with those, those criterion of high-quality HCBS, but in the current form, we acknowledge that, we have work to do here, to provide a better clarification, about that.

CHAIR NERENZ: Well, thank you. That -- that's very helpful. And, and, certainly, you know, understand, I appreciate the difficulty that you face, because you're taking something from a non-medical-care-environment, you're trying to bring it into an environment, by which, quality measures, definitions, concepts, are largely set in the medical care context. I, I fully appreciate the difficulty.

And, I think, what, you know, one of the things you said, a little while ago, if there exists a consensus, in this domain, about the characteristics of a good program, say, patient-centeredness is a concept, what would have been more convincing, to me, about the validity of a measure, now of that, would be whatever body of work was done, to lead to that decision, to say this is a characteristic of a high-quality program.

There must have been some previous thought, previous empirical data, I would've found that more convincing, than this pattern of inter-correlations.

Because, in fact, if you bring forward 14 independent measures, and it turns out, empirically, they're utterly independent, actually, that doesn't bother me.

Each one of them could be perfectly valid and not -- and related in any way to each other. So anyway, I -- we're kicking a dead horse, here, now.

DR. STOPLE: This is another very tricky area, too, and admittedly, a fairly nascent measurement space, so the committee that reviewed these indicators was, they were looking through an environmental scan, for measures of person-centered planning, just called attention to the fact that there's an absolute dearth of measures, for HCBS.

And so it's, this is -- it may not feel like it, but it's pioneering work and it, this, this was pioneering work 25 years ago, so

I'm very happy to see that, this come forward.

I would -- I also want to thank you, Dr. Li, for acknowledging the challenges. That -- that candor, actually, goes a long way. Some people get overly defensive on these discussions, but I appreciate you talking with the group, candidly.

But, one other area where, at least, from the staff perspective, has been, for some of the SMP Members, where we, we thought this submission could have some improvement, was just, I'm the, how -- the way that you reported it, the, the data element validity.

We had just a list of citations. And it was clear that there's been a lot of work that's been done, but a lot of it wasn't, it wasn't well-summarized, in this submission.

Is there -- is there parts, about the data element validity that you could speak to that would help the, the Panel understand, what has been conducted, in the past, and, and what those results were that point to the validity of the instrument of itself?

DR. LI: Yes. There has been a lot of work done, in the past, to corroborate the, the measures that we're putting forward, with real life significance.

So the, the table you're showing, right now, actually, is a great summary of that. Basically, we, we identified some measures that we put forward and located cases, where states, independently, the, the state's independent usage of those measures and, or they, they have some kind of equivalent measure, so that those two can be corroborated.

So the, the job goal one, you, you can see that the, the state population living, in urban areas, is not something that we collect, as part of NCI, and it's just an external source and we, we found it to be correlated -- and I note, this, again, goes back to Dr. Nerenz's point of, whether this is a confounding fact, or is this actually demonstrating that we have a good measure here.

But, I suppose we can, we can have this discussion, later, in a more focused way, and moving forward, actually, if there, there are some other cases in the issue for and response for that we provided, we noted that, several states, like Arizona, Massachusetts, and Kentucky, each, used their own way of corroborating.

For example, Arizona looked at employment and they, they worked on provider rates and show that they are, actually, important in facilitating District employments.

And, in Massachusetts, the State Department of Developmental Services actually corroborated our measures of Social Connectedness, which is the loneliness measure, with their own state-level data source, for using their licensure and certification data.

So again, this demonstrates that we

are not measuring something that's conjured up, or without basis. In Kentucky, there's a, there's a relation between the measure that we provided on job goal and social connectedness, with their, their own division efforts in those areas.

So I know, we shouldn't dwell on the same information that we submitted, but that, that's our, kind of, our summary of some of the existing work that has been done, to show that there are validity, external validity evidence to the measures that we put forward.

CHAIR NERENZ: I -- understood. Thanks, Dr. Li. Dr. O'Brien, then Dr. Romano.

MR. O'BRIEN: Yes, the comments that I provided didn't make it into the discussion guides. I don't know, if it's appropriate to raise them, now, or not, or --

CHAIR NERENZ: Oh, you can --

MR. O'BRIEN: -- or maybe I could --

(Simultaneous speaking.)

MR. O'BRIEN: -- raise them and not expect --

DR. STOPLE: Sorry, if we --

MR. O'BRIEN: -- if I -- I don't have --

DR. STOPLE: -- so please, go ahead.

(Simultaneous speaking.)

MR. O'BRIEN: This could just be something to, kind of, think about, for a future submission, if you're not prepared to address them, now, but there are three things.

The first thing was about the case mix adjustment, for two of the measures, and you provided calibration results, in the Excel tab, it was Tab 2b3.8.

And, I was confused, because when I saw those results, it looked like it was illustrating large discrepancies, between observed and predicted values, across deciles and predictive risk.

And, ordinarily, we'd go on, we'd go

on to expect to see much closer agreements on some of the deciles, or it's, either, kind of, two-fold difference, between observed and predicted.

So I can rattle off my, my three comments, and let you reply, or not, or I could go one-by-one, should I keep going?

DR. STOPLE: It's up to you, on how you'd like to proceed. If you'd, if you'd like the developer to respond to the first one, it might be a little bit easier.

MR. O'BRIEN: Sure. Yes, if you're prepared?

DR. LI: Sure. I'm assuming you're talking about the, the Excel sheet Tab 2b3.8, where we listed the life decisions scale and the community inclusion scale deciles --

MR. O'BRIEN: Okay.

DR. LI: -- does that -- yes that's, that's the table --

MR. O'BRIEN: Yes.

DR. LI: Okay, perfect. So I'm assuming, you're also highlighting the Decile 2, in the CC-4, with predicted to observed ratio of, almost, 1.94.

So this has to do, in my view, with the, the way that the scales are created. So first, I want to preface this, by saying, I don't have the full answer to this, because it, it will require us to dig deeper into the, the technical reasons behind some of those, the calculation of the numbers.

And I will devote time to do that and present that in the, the final submission. But, just based on my understanding of how the scales are created, I think I can provide some perspective on that.

So basically, the life decision scale are made of several factors that are actually stand-alone instrument items. So I'm pulling over -- I'm pulling out the, the actual syntax that creates those.

Basically, if it's three items, you have to have, at least, two valid scores to enter the calculation for the, for the scale measure. Because, otherwise, you are just reporting on one item and that's not a scale.

So what we, what we have to operationalize was, was that, we had to make a compromise and, and look at, by increasing the threshold of valid items required, we're, we're losing valid responses in, in return.

So it's a tradeoff between more inclusive -- more inclusion, or, or, or better available -- or better availability of the score, so that, as you can see, that has an effect on the deciles of those scores.

So if you'll look at the distribution of the score, from, from low to high, it, the lowest, the lowest score, it starts at .36.

Oh, I mean, the -- there is the, the mean observed scale at .00. But the -- because the, the model skews -- the model comes out so skewed, you're looking at a predictive scale value of 3.636, already, from Decile 1.

So this -- and, and you can see, from the observed scale score that, it's not a continuous distribution, either. It -- there has -- there are cutoffs in -- not cutoffs, but discontinuous patterns, in here. You can clearly see the quarter, the third, the half, the three-quarters, and so on.

So my long response was to say that, due to the nature of the, the scale, a creation, there is a mismatch between the predictive score, the scale score and the observed score, and that's why we're observing much, much higher observe ratio, at the lower end, and, and better, better ratios in the middle, and that is -- that is, actually, good.

Because, with a lot of those -- with a lot of our measures that we're submitting, we want to avoid the ceiling effect, because when we're effect -- when we're measuring a state's performance, we don't want them to all say, everything is doing well, we're, we're nearing 100 percent. That way, we get no discriminative power.

So we really need something that's in the -- that's near the middle and that's why you are seeing that, in the middle, it tends to perform better, and near the lower end, is where it's really problematic with predicted scale scores, just because, the model is so skewed.

Sorry, I don't know, if it answers any of your, your questions, sufficiently, like I said --

MR. O'BRIEN: Yes, thank you. It --

DR. LI: -- I --

MR. O'BRIEN: -- it -- it sounds like I, probably, misinterpreted it, a little bit, of what's in the Excel file.

I, I guess, I would've expected that, once the scores are aggregated, over a large number of patients and providers that those -- a discreteness, you're pointing to, and, like, the column for the mean observe scores would have, somehow, been averaged out and would've had more, more of a continuous measure.

But I, I didn't notice that they were multiples of things that were, you know, denominators of three and four and, and, and whatnot, so I, I think I misinterpreted. But, thanks.

The other comment was, basically, just based on the idea that each state is doing their own survey, and so presumably, they've got different interviewers.

And, a lot of the data you prevented, presented suggested that, there's imperfect agreement between raters for, sometimes, on the, you know, the, if the same survey was assessed, by two different raters, you might get two different answers, so imperfect repeatability, between interviewers.

So it raises the question, if you have

different interviewers, operating at different states, should some of the differences that you, you might consider be, kind of, random, imperfect reliability that could cause systematic differences, across states?

(Simultaneous speaking.)

DR. LI: Yes.

MR. O'BRIEN: And the states, they're also differing in their, kind of, their response rates and rates of valid responses and other things that, kinds of, differ, you know, things that might be considered random, are somehow systematic, once you're, kind of, comparing some results, from two different states.

DR. LI: Yes. Thanks -- thanks for the, for that note, because it is actually something we are dealing with, constantly, in our actual surveying, because one part of what we do is to make sure that we're working with so many different states that, everybody's operating on those, on the same terms.

And that's why we have the consistent training going on, throughout the year. So our survey starts July 1st, of every year, and goes into the next year, June 30th. That's the same very every state.

So we, we, we try to have the states field the surveys around the same time. They, they -- there are different approaches of hiring the, the actual surveyors, but they all have to pass standardized training.

And, when they -- and we have a, a central process of having our lead trainer, Dr. Giordano, training the lead trainers, within each state.

And, and after the -- after the states gets a chance to be, to, to have their surveyors trained, they then send out the, the surveyors. They're, they're not going out empty-handed, they, they all have the standardized tools, they all received the, the same guidance on how to treat questions, regardless of which state they're from.

We have trained surveyors, from Hawaii, and, and, from the northeast, they all, they're all operating on the same terms, and there are shadowing that's being carried out. I think that's what you're referring to, as to raters having an agreement.

MR. O'BRIEN: Yes, I, I'm blanking on which analysis it was. But, I didn't go back and refresh my memory, just, just now, but it -- they were, they were, I thought they were --

DR. LI: We did --

MR. O'BRIEN: -- there was an exercise you did that involved two interviewers, you know, re-interviewing.

(Simultaneous speaking.)

DR. LI: Yes. There, there are -- there have been a formal, shadowing studies that we did, with states that are across a wide, geographic range, where we would collect answers, on two different raters, one having the gold standard, which is typically the, the most senior trainer in that, in those scenarios, versus a relatively newer trainer -- a relatively newer surveyor, I should say, and compare their scores.

And, if I remember correctly, all of the, the shadowing studies that we did showed a medium to high agreements on, on their agreement levels. And then, there are more informal --

(Simultaneous speaking.)

DR. LI: -- more spontaneous --

(Simultaneous speaking.)

DR. LI: -- shadowing.

(Simultaneous speaking.)

DR. LI: Sorry.

(Simultaneous speaking.)

(Audio interference.)

DR. STOPLE: Sorry. Sorry to interrupt you. Let's, let's -- I'm really cognizant of the time. We're actually bumping up against the, the end of our meeting, so can we just summarize this last point, and then, let's go to --

DR. LI: Yes.

(Simultaneous speaking.)

DR. LI: Sure. So there are processes, standardized processes that are created and maintained, by the third party, which is us, going into, into different states, in a standardized way, and that's how we address issues, like, variations across states.

DR. STOPLE: Thanks, Dr. Li. I, I apologize for jumping in. Dr. Romano, you've got the --

DR. LI: Yes.

DR. STOPLE: -- the last word here, before we move forward with the vote.

(Simultaneous speaking.)

MEMBER ROMANO: Well, this is where I really wish that Sherrie Kaplan were here, to put a spade in this discussion. Because, I think, the fundamental problem -- this is a wonderful measurement program.

It's -- it's very important that these surveys are done that these measures are collected and that there's systematic benchmarking and so forth.

The question is, what are we voting on, as a quality measure that is, you know, consistent with NQF standards? And I would argue that, you know, when you have a, a set of measures, like this, and even the title, 3622, is plural, measures.

When we have a set of measures, we have to figure out, what are the measures, right, that are actually being used, for accountability purposes?

And so I, I would argue that, what I'm reading, from this, is that, it's really the scales --

(Audio interference.)

MEMBER ROMANO: Sorry. And it's really the scales. Specifically, for example, the Community Inclusion Scale and the Life Decision Scale that are really important.

So under reliability, for example, you report the Cronbach's alpha, for those scales. You report, in validity, how those scales, specifically, the Life Decision Scale and the Community Inclusion Scale, were carefully risk-adjusted and how they were carefully used to assess states performing above, or below, the benchmark.

You report in, in fairly strong detail, about the construction of these scales. And so most of us, in this field, would consider these scales, to be the measure.

We don't -- we shouldn't need to vote on the individual components of the scale, if we agree that the scale hangs together and the scale is, fundamentally, what is being used, as an accountability measure.

So I -- I'll, I'll just put that -- I don't know how that affects our voting, here, today, but I, I would suggest that, it would just be a lot clearer, if we were voting on the scales, rather than, voting on 14 separate items that go into the scales.

DR. STOPLE: Noted. Of course, we can't change our process, in the middle of the meeting, but, but something that we'll need to think through, carefully. So thanks, for that point, Patrick.

We had one other comment that wanted to be made. Is it, Alixe Bonardi, did you want to jump in?

MS. BONARDI: Sorry, I didn't hear it. Sam, did you call on me?

DR. STOPLE: Yes, I did. Go ahead.

MS. BONARDI: Oh, thank you, thank you. Hi, this is Alixe Bonardi. I, I just wanted to, to make one note that, in, in response to some of the comments and, and I appreciate the comments, about how this is a challenging thing to, to bring what is a home and community-based services, really, into a quality monitoring that's more, from a medical framing.

I -- I think that, in the development of this content, we, we did not really build out a lot of the quality framework, in this context, recognizing that that had already been done, by NQF, in the context of the HCBS Measures of Quality Report that Dr. Li referenced, back from 2016, which, both, laid out domains of quality and, also, gap areas.

So recognizing that this is a rather nascent area of measure development, certainly, home and community-based services, I, I just wanted to put forward that, the approach we've taken is to take what is a body of work that has developed measures and, as Dr. Li pointed out, has certain measures of, certainly, reliability and construct validity that could be put forward, for NQF to consider, as a, as a pretty efficient way, to begin to fill some of the gaps in quality measurement that really is, is -- that's our intent here, and that's -- that's all I wanted to lay out.

Because, I, I, I do think there are frame, there is a framing of quality measures, but, perhaps, I think, we, we may not have laid it out, in as much detail, as we had hoped.

DR. STOPLE: Thanks, very much, Alixe. Okay, given our time limitations, I think we need to move forward, with proposing an approach for the -- as you know, our process dictates that we would need to have votes on the individual measures, if desired.

Now, what we did, with the CAHPS measures is -- we didn't actually tease any of those out. So for HCAHPS, for example, has, you know, I think, nine, or ten, different domains that are considered separately, as separate measures.

We may elect to feel those out, but we would default to voting on them, as a group. So if they're -- if there's any one component that any Member of the Panel wishes to pull out to vote on, separately, I would suggest that we move forward, with inviting you to identify what those individual components might be.

CHAIR NERENZ: And, Sam, just response and friendly amendment, I think, you know, the time we have in front of us really only allows sort of, one vote. If we start peeling out measures and then, we'd have to discuss each one, we'd have to individually vote it. There just simply is not time in the window.

My suggestion is, why don't you, essentially, call the question on this one, thing. But -- but, my strong request, to the staff, would be, take into account some of the comments, we've all made, about the, the challenge of trying to fit, essentially, round peg into square hole, here.

And how it, it may be that, something like Patrick's suggestion about focusing on scales is something that, you know, perhaps, could be attended to and then, we'd have a fewer number of units to, to test.

I just feel very badly, about how we treat this, no matter what we do. I think, if we vote and pass it through, with all the concerns that have been raised that's not good.

If we, essentially, vote to fail it that's not good. But -- but, I think, the time -- it -- if we need a formal vote on something, all we can do is, is treat it, as a package, I, I don't see another way to do it, right now.

DR. STOPLE: All right, very good, Dave. And, you're probably right, we definitely need to think through, how to do this economically and to still pay the appropriate respect to our process.

So with that being said, then, it sounds like, Dave, the motion you're putting forward is for us to treat this, as a package, and then, to vote on validity from here, is that correct?

CHAIR NERENZ: I, I -- I mean, it seems almost inevitable, looking at the clock. I don't know how else to do it.

DR. STOPLE: Okay.

CHAIR NERENZ: I mean, I don't know if there's any, sort of, formal process for tabling, but that doesn't mean, just put it to tomorrow. That would just mean, like -- I don't know what it means. I -- it's never been an option we've had, before.

DR. STOPLE: Yes, it's not one that we've, we've exercised, in the past. It's -- this is the body of voting that we need to take and that we need to address and we get the meeting and we need to do it, or reschedule more meetings.

DR. MA: I don't know, if, or how, Sam would do -- prepare -- would -- have had prepared a one question, for the subgroup, do you want to vote them, as one package, or individually, but

I think, I don't hear any -- you know --

DR. STOPLE: And it --

DR. MA: -- anyone is against voting, as a package, you can speak up, otherwise, we can move forward with package vote. We prepared three different ways for you to vote.

(Simultaneous speaking.)

DR. STOPLE: Yes, I'm not hearing any, sighs, so let's go ahead and move forward with the package vote.

MS. INGBER: Okay. Voting is now open, for validity, on Measure 3622. Your options are high, moderate, low, or insufficient.

(Voting.)

MS. INGBER: Okay, we have all the votes in. Just give me a moment. All right. Oh, thank you, Caitlyn. So I think, we'll see, we have zero votes -- oh, sorry -- for Measure 3622, on validity.

We had zero votes, for high. Four votes, for moderate. Zero votes, for low. And, three votes for insufficient. Therefore, consensus was not reached on this measure.

DR. STOPLE: Okay.

CHAIR NERENZ: That captures the essence, pretty perfectly. That's not bad.

DR. STOPLE: Well, Dave, perhaps, it will be helpful, just before we move on, I recognize, we're really short on time, for you to summarize a couple of advice points, for the developer, in helping them, as they're preparing to go forward with this conversation to the full committee.

CHAIR NERENZ: Well, I guess, one thing, if I, if I understand process, our consensus not reached, means that the measure can move on to a standing committee and that's important to note.

We didn't just torpedo and fail this measure. But, obviously, you know, the transcript will capture the concerns and I don't want to try to reiterate all those, here.

I think the suggestions that came up were, largely, some things that might be done, it, in, yet, a future cycle, assuming the, you know, goes through this standing committee, in a, in a positive way.

Or, I don't know, if there's any opportunity to do some additional things, prior to the standing committee that weren't shown to us. Again, that's a process thing, I don't know.

So again, not to take up any more of our time. I think the comments have been great. I think the members of our subgroup, who really put --

(Audio interference.)

CHAIR NERENZ: -- a lot of time into thinking about this. I thank the developers, again, really, a heroic job of, of trying to make this fit in a, largely, unfamiliar process.

And, you know, let's just let it run from here. Offline, we can consult with staff, about exactly how to do it, I just don't think we ought to take more time, with the whole group, right now, because I --

DR. STOPLE: I -- very good. Thanks, Dave, appreciate it. And, a very big thanks, to the developers and the, the Panel, for this discussion. Sai, back to you.

(Simultaneous speaking.)

DR. MA: Yes. Thank you, everyone. I'm apologizing, we are run over time, but, before we open up for public comment, I just want to remind every developer that, NQF does provide technical assistance.

The last measure is a perfect example, where we can provide technical assistance, to help you get through the next phase, with the standing committee.

So, Sam, is your guy, reach out to him. And, now, we are open for public commenting.

(Pause.)

DR. MA: We do have, had a lot of

comments, in the chat box. They will be saved and put on the record. So I'll wait, for another few seconds, to see if we have any public comments.

(Pause.)

DR. MA: Okay, hearing none, I think, we can close out, for today. It's been a really long day. Thank you, so much, for everyone's attention and participation and the robust discussion.

I know, I have learned a lot. I hope everyone feels the same way. And, I'm not going to take more time from you, all, we'll see everybody, tomorrow, at 11:00 a.m., Eastern Time. Good night.

(Whereupon, the meeting in the above-entitled matter was concluded at 3:55 p.m.)