**NATIONAL QUALITY FORUM**

**+ + + + +**

**SCIENTIFIC METHODS PANEL**

**SPRING 2020 MEETING**

**+ + + + +**

**TUESDAY**

**APRIL 1, 2020**

**+ + + + +**

**The Panel met via teleconference, at 9:00 a.m., Dave Cella and Dave Nerenz, Co-Chairs, presiding.**

**PRESENT:**

**DAVE CELLA, PhD, Co-Chair**
**DAVE NERENZ, PhD, Co-Chair**
**J. MATT AUSTIN, PhD**
**BIJAN BORAH, MSc, PhD**
**JOHN BOTT, MBA, MSSW**
**DANIEL DEUTSCHER, PT, PhD**
**LACY FABIAN, PhD**
**MARYBETH FARQUHAR, PhD, MSN, RN**
**JEFFREY GEPPERT, EdM, JD**
**LAURENT GLANCE, MD**
**JOSEPH HYDER, MD**
**SHERRIE KAPLAN, PhD, MPH**
**JOSEPH KUNISCH, PhD, RN-BC, CPHQ**
**PAUL KURLANSKY, MD**
**ZHENQIU LIN, PhD**
**JACK NEEDLEMAN, PhD**
**EUGENE NUCCIO, PhD**
**SEAN O'BRIEN, PhD**
**JENNIFER PERLOFF, PhD**
**PATRICK ROMANO, MD, MPH**
**SAM SIMON, PhD**

**ALEX SOX-HARRIS, PhD, MS**

**MICHAEL STOTO, PhD**

**CHRISTIE TEIGLAND, PhD**

**RONALD WALTERS, MD, MBA, MHA, MS**

**TERRI WARHOLAK, PhD, RPh, CPHQ, FAPhA**

**ERIC WEINHANDL, PhD, MS**

**SUSAN WHITE, PhD, RHIA, CHDA**

**NQF STAFF:**
**SHANTANU AGRAWAL, MD, MPhil, President and CEO**
**ASHLIE WILBON, MS, MPH, FNP-C**
**SAM STOLPE, PharmD, MPH**
**MIKE DiVECCHIA, PMP**
**HANNAH INGBER, MPH**
**CAITLIN FLOUTON, MS**

**ALSO PRESENT:**
**KATIE BALESTRACCI, Yale CORE**
**LISA BERGERSEN, Boston Children's Hospital**
**LISA SUTER, Yale CORE**

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

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

(9:06 a.m.)

MS. WILBON: Good morning, everyone. This is Ashlie Wilbon, the technical expert from NQF. I want to thank everyone for joining us today. This is NQF Scientific Methods Panel Spring 2020 Evaluation Meeting.

We're certainly happy to have our Methods Panel members on as well as those of you who have joined and members of the public and our developers given these trying and special circumstances we're all living through at this point.

So, we're happy to have you and we are looking forward to an interesting discussion today despite things that may be going on around us. Again, welcome.

I did want to hand it over to our CEO here at NQF, Shantanu Agrawal, to give us some welcoming remarks, Shantanu.

MR. AGRAWAL: Thanks, Ashlie. Can you hear me?

MS. WILBON: Yes, we can hear you.

MR. AGRAWAL: Excellent. I'll just do a quick welcome -- very unique and trying times. And so the fact that this committee would take the time to be part of this meeting and, more than that, be part of the work, we very much appreciate.

I think we are trying to move forward on our endorsement work as best as possible. We will likely be giving some guidance to measure developers and giving them some leeway in the coming weeks in order to accommodate some of the bandwidth constraints that developers and others are feeling right now.

But again, we've gotten a host of measures for this endorsement cycle which I think is a great show of support. And just a great statement on the part of our stakeholders.

We internally have reviewed the agenda for today and there are a host of very interesting, I think, measurement issues. And so I think you'll have a really robust dialogue

throughout the day.

And I think you'll be able to provide a lot of guidance and recommendations back to us so that we ensure that our approach to measure endorsement remains rigorous and we are tackling some of these new issues as they come up.

So, again thank you very much. I want to thank all of you for attending and want to thank the NQF staff as well. They have really done yeoman's work to create a great agenda even in challenging times.

So thank you, Ashlie and team, and thank you all for attending. I'll turn it back over to Ashlie.

MS. WILBON: Thank you, Shantanu. And I did want to see is -- Kathleen, are you there? Kathleen Giblin is our acting senior vice president of quality measurement.

I wanted to check in to see if you would like to say any words or if you're there. If not, we'll give an opportunity to our co-chairs to provide some welcoming remarks and then

we'll get started here very shortly.

Okay. I don't think Kathleen is on. So, Dave and Dave, would you like to give some opening remarks?

CHAIR CELLA: This is, excuse me, this is Dave Cella. Good morning, everyone. Thanks for attending.

We have a lot scheduled to talk about and I won't delay any further so that we can be as efficient as possible and get right to some of the meat of the meeting. Thanks.

CHAIR NERENZ: Yes, Dave Nerenz. I'll echo that. Really tough time for folks and particularly those who are active clinicians, who are right on the front lines of what's going on around us.

Thanks for your ability to spend some time with us. So, without further ado, let's get to it.

MS. WILBON: Thank you. Thank you both. And just a couple of quick meeting and webinar reminders. You can see from our agenda

today we do have -- we'll be meeting in about two-hour blocks with a couple one-hour breaks.

We wanted to give folks sufficient time to be able to grab a bite to eat, stretch your legs and do other -- make sure you can check in with your other obligations throughout the course of the day.

We do understand that you guys have a lot of other things going on around you. So, if you need to step away outside of a break, please feel free to do so.

We do just ask that if you are stepping away, if you try to be present during our measure discussions, particularly if your subgroup is up for discussion, then we ask that you try to attend during those times so that we can make sure that we are able to maintain a voting quorum for those measures.

And please feel free to use the chat feature in the webinar. Our team is on and will be able to provide any support you need.

We'll do our best to mitigate any

technical issues you might have in interacting with us today. And we found the chat box very useful to communicate during meetings. So please feel free to use that.

As we go, since we're not sitting in a room together and we don't have our table tents, our proxy for that via the webinar is the raise your hand feature. You each should be able to raise your hand if you would like to speak at any point during the discussion.

NQF staff and the co-chairs will be watching those indicators to see who would like to raise their hand and we'll do our best to call on folks in a timely and orderly way. So, we'll try to catch people in order of when they raise their hand to make sure that everyone has an opportunity to speak.

And if we're not catching you, please feel free to send us a chat or kind of summarize your comment in the chat and we'll make sure to interject that comment so that we can call on you accordingly.

We are transcribing the discussion today similar to how we would do for an in person meeting. We do have a court reporter who is dialed in so that he can get an accurate transcription and recording of today's discussions.

So we do have a couple of asks in order to make sure that goes smoothly. Particularly if you're a presenter, if you could, and that would be essentially anyone who is a member of the Methods Panel, if you could remember to keep your line muted on your phone so that we don't get the background noise and echoing during the discussion that allows him, the transcriptionist to hear the speaking clearly and make sure that he's able to get a clear recording.

We will also ask that if you can remember to be sure to introduce yourself as you're speaking. Again, it helps him keep track of who is making which remarks.

You may find us interrupting if we

don't know who is speaking to make sure that we have a good idea of who is contributing to the discussion. One more note that we were given as a reminder so that we can make sure that the line is as clear as possible, if you are able to dial in and use a set of headphones or a headset of some sort to speak rather than having on speakerphone, I think that does also help with background noise and the clarity of the sound.

So, if possible and you're able to do that please do. And again, we'll be doing our best to mitigate any technical issues as we go along.

I think everyone is familiar with our team at this point. I think we've all sent enough emails at this point that everyone is pretty familiar.

But I did want to just take an opportunity to recognize my team. We've all worked really hard.

They, in particular, have worked really hard to pull this together and transition

and pivot from an in-person meeting to conducting the meeting over the web and they've done a really great job at being flexible and accommodating for folks.

So, thanks to the team. I did also want to just recognize Sam Stolpe, who I think you all met last cycle. He is a senior director here.

He will be my backup today in case anything comes up and he'll be helping with some of the measure discussion later on this afternoon.

I did want to just offer an opportunity for each of the team members to do a brief introduction and to say hello to everyone this morning.

Mike, are you there? You want to say hello?

MR. DIVECCHIA: Hi. Good morning, everybody. This is Mike DiVecchia and I am the project manager for the SMP. And I'm looking forward to the next couple of days.

MS. INGBER: Good morning, everyone. This is Hannah Ingber. I'm also very excited to hear the discussion. Thank you all for joining us.

MS. FLOUTON: Good morning. This is Caitlin Flouton. I am a new analyst here. I am available in the chat and in email. If you need any help, feel free to reach out and thanks for joining us.

MS. WILBON: Thanks, team. So, at this point in time we'll get a sense of who from the Panel is on the line. We will be combining the roll call with our disclosure of interests.

If you recall, any time we do measure review, we will need to do a disclosure, an oral public disclosure of interest. And so bear with me.

Some of this process is scripted and it will be very familiar to you. But it is important that we do this and make sure that we have any oral disclosures that are needed.

I will point out that I know a couple

of our West Coast Panel Members may be dialing in a little bit later and we'll just make sure to follow up with them as we get a little deeper into the discussion.

So, we'll start with our co-chairs. I think they are both on at this point. Dave Nerenz.

CHAIR NERENZ: Good morning. Dave Nerenz here. No conflicts to disclose.

MS. WILBON: Apologies, Dave. Sorry, I should read my little script here. So, you each received a disclosure of interest form from us before you -- as you were named to the Panel.

And it asked you about some general professional activities. And then we also provided you with a measure-specific disclosure of interest.

It's specific to the measures we received and will be under review for this cycle. So, in the interest of transparency today we're going to ask you to orally disclose any information you provided that you believe is

relevant to the discussions today and the matters that will be under review.

We're specifically interested in grants research, consulting, measure development activities related to the work today. There are seven measures for review today on the agenda.

So, we'll be specifically -- looking to specifically hear any conflicts related to those. And again, most of you have heard this before. But just a couple reminders.

You sit on this group as an individual and you do not represent the interests of your employer or anyone who may have nominated you for the Panel. We're interested in disclosures of both paid and unpaid activities relevant to the work today.

And just because you disclose does not mean you have a conflict of interest and again, the oral disclosures are in the spirit of openness and transparency.

And just another brief reminder that if at any point you believe that you may have a

conflict of interest or someone else on the Panel may have a conflict we ask that you let staff know so that we can address it in as real time as possible or as soon as possible.

And at this time we'll jump back in. Apologies for that. Dave Cella?

CHAIR CELLA: Good morning, again. This is Dave Cella. I am an unpaid member of a board of directors of the PROMIS Health Organization which is a non-profit foundation that supports PROMIS.

And PROMIS will come up in one of the discussions today or tomorrow.

MS. WILBON: Thanks, Dave. Matt Austin?

MEMBER AUSTIN: Yes, good morning. I have no disclosures to offer related to this --- this morning.

MS. WILBON: Thanks, Matt. Bijan Borah?

MEMBER BORAH: Hi, good morning. I have nothing to disclose relating to the measures

we will be discussing today.

MS. WILBON: Thank you, Bijan. John Bott, are you there?

MEMBER BOTT: Yes, hello. Good morning. John Bott, I have no disclosures of interest related to the work of the Panel. Thank you.

MS. WILBON: Thank you, John. Daniel Deutscher?

MEMBER DEUTSCHER: Yes, hello. I have no conflicts or disclosures related to any of the measures submitted during this cycle, thanks.

MS. WILBON: Thank you, Daniel. And I just want to recognize Daniel who is calling in from Israel, I believe. So, thank you again for your engagement. I appreciate that. Lacy Fabian?

MEMBER FABIAN: Good morning. I have nothing additional than what was disclosed on the forms for the post-acute care measures. Thank you.

MS. WILBON: Thank you. Marybeth

Farquhar?

MEMBER FARQUHAR: Good morning. This is Marybeth Farquhar. I have nothing to disclose with relation to the measures.

MS. WILBON: Jerry, I'm sorry, Jeffrey Geppert?

MEMBER GEPPERT: Hi, good morning. Nothing to disclose for today.

MS. WILBON: Thanks, Jeff. Larry Glance? Okay, I think Larry is --

MEMBER GLANCE: Can you hear me now?

CHAIR CELLA: Yes.

MS. WILBON: Hi, Larry, yes.

MEMBER GLANCE: Hi, good morning. Sorry about that. I was muted. I don't have any additional things to disclose, thank you.

MS. WILBON: Okay, great. Thank you. Joseph Hyder? Joe, are you there? Okay, Sherrie Kaplan?

MEMBER KAPLAN: Hi, I'm a consultant to the AHRQ-funded Shared Decision Making Measures Development group. But that's not

relevant to the measures today.

And the PCORI, I have I -- I am a PI on a PCORI-funded child-reported -- patient-reported outcome, Type I diabetes, but also not relevant for measures today.

MS. WILBON: Thank you, Sherrie. Joe Kunisch?

MEMBER KUNISCH: Hi, good morning. This is Joe Kunisch and I have no conflicts of interest to report.

MS. WILBON: Thank you. Paul Kurlansky? I think Paul might have been having some trouble getting in. Paul, are you there? Maybe on mute.

PARTICIPANT: I know he said that he could not make both days. This may have been the day he can't make.

MS. WILBON: Yes. I think we had some communication from him this morning. He was trying to get -- well, we'll come back to Paul. Zhenqiu Liu?

MEMBER LIU: Yes, hi. I'm involved

with two measures we discuss today and tomorrow. One is 3559. One is 2539.

MS. WILBON: Thank you, Zhenqiu, appreciate that. Jack Needleman, are you there?

MEMBER NEEDLEMAN: Good morning. Nothing to disclose.

MS. WILBON: Thanks. And Jack is also joining us from the West Coast so thanks for dialing in early, Jack. We appreciate that. Gene Nuccio?

MEMBER NUCCIO: Hi, yes. Hi, no disclosures on any of the measures today, thank you.

MS. WILBON: Thanks, Gene. Sean O'Brien? Jen Perloff?

MEMBER PERLOFF: Hi. No conflicts of interest.

MS. WILBON: Thank you. Patrick Romano? I know Patrick is one of our West Coast folks so he may be dialing in a little bit later. Sam Simon?

MEMBER SIMON: Good morning.

Mathematica is a measure developer and we have a contract with CMS to develop a measure that is very similar to 3559.

It's basically the clinician level analog to 3559. So, that's the only thing I have to disclose.

MS. WILBON: Thank you, Sam. Alex Sox-Harris?

MEMBER SOX-HARRIS: Good morning. I'm an unpaid workgroup member for the American Academy of Orthopedic Surgery for some of their measure development efforts. But nothing related to the measures today.

MS. WILBON: Okay, thank you. Michael Stoto?

MEMBER STOTO: Good morning, everyone. I have nothing to disclose for today's meeting or tomorrow's.

MS. WILBON: Thank you. Christie Teigland?

MEMBER TEIGLAND: Hi, good morning. And I also have nothing to disclose related to

the measures we'll be discussing over these two days. Thank you.

MS. WILBON: Thank you. Ron Walters?

MEMBER WALTERS: Good morning. I have nothing to disclose relevant to the discussion of this committee.

MS. WILBON: Thank you. Terri Warholak?

MEMBER WARHOLAK: Good morning. I have nothing to disclose related to the measures today. And as you will hear, there's a little background noise but I'm actually in my apartment.

So, I'm going to try to keep myself muted as much as possible.

MS. WILBON: Okay. Thanks, Terri, appreciate that. Eric Weinhandl?

MEMBER WEINHANDL: Good morning. This is Eric. I am an employee of Fresenius Medical Care North America.

It's a supplier of dialysis services and products and I have no conflicts of interest

with respect to the measure development. One of the measures being discussed tomorrow relates to emergency department readmissions for dialysis facility evaluation. So, my employer is certainly subject to that measure.

MS. WILBON: Okay, thank you. Susan White? I believe Susan let us know she probably would not be able to attend. But just checking in, in case. Susan, are you there?

Okay, she said she may be able to pop in throughout the two days. So, thank you everyone for that. As we get into the measure review later on this afternoon we'll make sure that those who speak are identified.

For those of you that are on the Panel, you'll note on the annotated agenda we have noted the conflicts that we've noted for each of the measures. So please be sure to review that, and if you are -- have been identified as someone who should be recused, please note that you should be refraining from discussion and voting on the measure. So, thank

you again for that.

Just a quick review of our agenda today. We're going to go through some evaluation updates, a process overview, some evaluation guidance discussion which is designed as the emailed you guys later last week to really give the committee or give the Panel an opportunity to discuss some of the overarching issues that were identified through the evaluation of the measures this cycle.

And we'll be trying to use that discussion to establish some common principles that we'll be keeping in mind as we evaluate the measures later on in the afternoon and early part of tomorrow.

And so we'll pick up with the measure evaluations later this afternoon. And tomorrow we will be spending the morning on measure evaluations.

And as we noted from our email update yesterday or a couple days ago, we will be kind of truncating the meeting on Day 2 to conclude

after measure evaluations, just keeping in mind the many external activities and things going on around us.

We'll try to keep Day 2 as brief as possible. But obviously, we'll need to make sure that we get through the measure evaluations for the cycle.

So, the criteria recommendations and evaluation guidance discussions will be moved to a future webinar for the Panel in the spring. So, thanks for that.

And we just went over the agenda. I did just want to give a brief overview of the meeting materials that we provided because we will be referring to many of these today.

Obviously, the agenda provides some good information about who we're expecting to attend, who should be voting on the various measures and what members should be recused from discussion.

The discussion guide will be a reference tool when we begin measure review. It

has all the important information about the measures that will be under review.

It also includes links to the measure information form and the testing attachments, as well as the additional information that was provided by developers in response to your preliminary analyses.

There are also several background materials that we provided, some for reference and some that we were hoping you would be able to review, to provide you with a little bit of context of some of the upcoming discussions.

We'll be referencing various elements from the 2011 NQF Testing Task Force Report as we start thinking about some of the evaluation guidance and policies and recommendations for changing our criteria and evaluation policies coming up.

We also provided the SMP Measure Evaluation Guidance document that we have created since the SMP has been in existence, which basically outlines for developers and others some

of the preferences that the Methods Panel has expressed in starting to review measures so that developers have a sense of how we might like to see data presented and so forth.

We also wanted to share the risk adjustment paper that Larry led and has just published and was supported by many of the SMP members as background as well as a paper from He, et al., on the PIUR which references a method for reliability testing that would be for several of the measures this cycle.

So, to just jump right in here and give everyone some foundation on where we landed with the measures this cycle. So, we had a total of 50 measures that were submitted to NQF for the 2020 cycle.

And about a little less than half of those were evaluated by the SMP. So, we had a total of 21 measures. And again, about half of those were new and the other half were maintenance measures.

We have three subgroups this cycle,

about nine to ten folks. And each group had seven measures.

We ended up with 14 that passed reliability and validity, two that were consensus not reached upon either reliability or validity, and five that did not pass reliability or validity.

And there are a total of seven that are slated for discussion today, which combines the two that were consensus not reached and the five that did not initially pass a preliminary analysis.

For most of those measures that did not pass, as well as the ones that were consensus not reached, the group did submit additional information for consideration. So, we'll be discussing that information in the context of reconsideration and submitting additional votes later on today.

And again, here is a breakdown of the type of measures that we got. We had a set of six top measures, one PRO-PM, one composite, and

the rest were outcome or intermediate clinical outcome.

I just want to pause there to see if there's any questions before we move on. But the next section is around updates on the Panel's, kind of, evaluation performance and some of the disposition of the measures that have gone through previous cycles.

But any questions about this cycle or meeting materials or anything like that before we move on? Okay, hearing none, I'll keep filing away.

So, we did want to kind of bring back to you some of the facts that we've been selecting since the Methods Panel has been in existence. And we've seen definitely some fluctuation in the number of measures that have been reviewed by the committee.

We certainly had a peak in spring of 2019 with almost 50 measures. And we've kind of come down in terms of the number of measures that have been reviewed from fall of 2019 to spring

2020.

And we've been more in the 20 range, which certainly seems a lot more palatable, I think, for us all. The number of measures that have -- where consensus was not reached, I think have certainly been decreasing over time.

And I think certainly the number is consistent with the total number of measures that have been reviewed by the committee.

And we'll actually share in the following slide a little bar graph that actually is able to show kind of how the committee is becoming more consistent in evaluating the measures and coming to consensus a lot more frequently since the committee has been working together.

So, this is what I was referring to. So you can see over time that the number of measures or the percentage of measures where consensus was not reached is trending down.

So that is certainly positive from our perspective, in that the committee or the Panel

seems to be coalescing around some common principles and tends to be, more often than not, on the same page about where the, you know, which measures are -- should be considered passing and considered for full endorsement and which should not.

I did just want to go back to last cycle to give a little bit more details of where we landed. We did, we evaluated 22 measures. We had 50 measures that were discussed at that meeting.

And that was a combination of measures being pulled by staff, being pulled by Panel Members, and also there were six measures where consensus was not initially reached in the preliminary evaluations. Ultimately, the SMP passed 16 of the measures and -- which was about 73 percent.

And there were four that did not pass, two that were eligible for Standing Committee revote and the -- both of those were pulled for Standing Committee -- by the Standing Committee

for reconsideration.

And there were also additional, those other two were pulled for further discussion as well. And we did have one measure that was withdrawn by the developer.

But if you recall, we did implement a new process starting last cycle where Standing Committees had an opportunity to pull measures for discussion if they had not passed the SMP and potentially be eligible for revote depending on the reason for it not passing.

And I think I saw a hand from Matt. Did you have a question or a comment?

MEMBER AUSTIN: I do have a question. But you may get to it in the future so I sort of inadvertently raised my hand.

But I'm curious to know if you have any specifics on measures that haven't passed that have come back to the Panel and if those have passed the second time or if you have any sort of statistics on those that we have reevaluated.

CHAIR CELLA: Ashlie, are you there?

MS. WILBON: Hi, sorry. I was talking on mute, apologies. It's a good question. I don't have those numbers in front of me but it's something that we can work on and bring back to a future meeting.

I think it's something we should be tracking. I do know this cycle that we did not have any that came back that were previously reviewed.

We have certainly, over previous cycles, seen more coming back. And typically, I think, this sort of experiential data, I can tell you that it typically does take about a year for them to come back.

So, they aren't typically coming back, let's say, in the following cycle. But, yes, it's a good question. I don't have it, but I can get it.

MEMBER AUSTIN: Thank you.

MS. WILBON: So, switching gears a little bit, but again just a bit of kind of Panel

maintenance if you will, we do have several members whose terms will be ending in the fall. And we did want to just do a check in with those folks.

We will be distributing a survey to you guys via email. And we just ask if you could complete that. It's literally one question to just let us know whether or not you would like to continue on and extend your term.

We're acting now because, depending on how many folks decide to let their terms expire we may need to do additional nominations and we need to do that before the next cycle. So, if you could complete that for us and let us know if at your earliest convenience that would be awesome.

I did also want to let folks know, Dave C., if it's okay for me to share that Dave will be, his term will be expiring in September and he unfortunately will not be returning with us. So, we will hang on to him for as long as we can through September.

But we'll be looking for another co-chair in September and we'll be in touch about that over the coming months. But we still have him for several months.

But I just wanted to give a brief update on that. Okay. So, we do have a few updates here on the process I wanted to share.

As I think most of you guys are familiar with this -- so I won't spend a lot of time on these. They are things that you have all seen before.

But I do think they are helpful reminders as we get into review, just to make sure we're all starting off on the same page. If I am going too fast or you need me to repeat something, please let me know and we will certainly go back.

But again, in the interest of time and not -- making sure we can get to the meat of what we're doing here as quickly as possible, I'll move quickly.

So, I did want to bring up the --

remind us about the process we implemented last cycle, which was around the ability for Standing Committee Members to be able to evaluate measures that had not passed the Methods Panel's evaluations.

So, similar to previous cycles, we have not changed the process, in that measures that passed the review or where the SMP does not reach consensus, those measures will automatically go to the Standing Committees.

And so, that has not changed. That's been consistent since we started the Methods Panel.

The newer process that we added last cycle was about consideration of measures that did not pass by the Methods Panel, by the Standing Committee and the fact that some measures may be eligible for revote if it's considered eligible.

And so, the next slide here talks in a little more detail about that. So, after the Methods Panel's final vote NQF staff basically

reconvenes with the co-chairs.

They review the measures that didn't pass, and discuss kind of the key issues on the rationale for why they didn't pass, and see whether or not they will be eligible for consideration by the Standing Committees.

And typically, measures that do not pass the SMP are not eligible for revote if it -- if the reason that it didn't pass hits one of these four bullets. So, if there was an inappropriate methodology or testing approach applied to demonstrate reliability and validity, there is incorrect calculations or formulas used for testing, the description or testing approach results, or data is insufficient for the Methods Panel to apply the criteria.

That would be inclusive of, you know, specifications being unclear or different information that would be needed to correctly calculate the measure. So, that kind of falls under that bullet or the appropriate level of testing was not provided or otherwise did not

meet our minimum evaluation requirements.

Typically, in our kind of screening of measures before they go to SMP we try to catch everything that does not have the right level of testing. But that's not always clear in our initial screening.

So, sometimes those do get through to the Methods Panel for review. So there is another stopgap at this point for us to be able to catch those measures.

I'll just pause there to see if there's any questions. I think we had some questions previously from the Methods Panel and other stakeholders about this.

So, I did want to just pause here and just sure that the Methods Panel requested that it was clear what the policies and procedure for that. Okay.

CHAIR CELLA: Just a sec. Jeff Geppert has a hand up.

MS. WILBON: Okay, Jeff, sorry. Hi.

MEMBER GEPPERT: Hi. Yes, so my

question isn't so much about the policy but where the, will there be a feedback loop, you know, from the Standing Committees back to the SMP about, especially -- I think in your earlier slide about 75 percent of the time there was agreement, about 25 percent of the time there wasn't.

So, it would be nice to know, you know, if there are some lessons learned from when they disagree.

MS. WILBON: Yes, that's a good question. I think for the last cycle, for spring 2019, when we first initiated a policy where they could pull measures for revote -- I do not have in front of me kind of the disposition of those measures, on whether or not they actually ended up passing, and what the committee's rationale was.

But it's certainly something we could do, we could bring back, to provide that feedback to you guys, certainly.

And I think even from an even broader

perspective, the measures that you do pass forward, those 75 percent, was there any kind of discrepancy in what the committee believed in terms of their recommendations for endorsement or for passing and not passing is certainly something we could bring forward.

I think we have stats on that. But I don't think I included it here. So, at this point it's just something we'll look into.

Okay, moving forward again to the ratings here. I think you guys are all familiar with this.

But I did just want to just recap a couple of things, particularly because I think the rating scale has been something that has been an element of discussion for the Panel in previous meetings, and to just recap a couple things, and differentiating high from moderate, for example.

Moderate is the highest rating that a measure can be eligible for if only data element testing was presented or if only face validity

was conducted. So, typically score level testing is required in order for a measure to receive a high rating.

And the rating for the measures should be -- should encompass your evaluation of, you know, whether or not the test was appropriate for the measure and the purpose of the testing, whether or not the scope of testing was adequate, so representative of the sample, representativeness of the population, the right sample size and then also of whether or not the results were acceptable.

So, I'm kind of thinking about those three elements and the findings or ratings all just to be considered. Again, here this is just a bit of a repeat.

Two-day consensus, we're looking for a quorum of 66 percent of the Panel Members. So, that will be a slide at the subgroup level. We will be loading just within the subgroups.

Unfortunately, we were not able to change the process at this point to do a full

Panel vote. It is something we're still considering.

So, for this cycle we will continue with our existing process for only subgroups to vote. I do see a hand from Sherrie. Do you have a question?

MEMBER KAPLAN: Yes. I have to be really careful with my phone because the off button is right above the mute button. And so, I don't want to turn the phone off.

But I had a question on the high/moderate. Can you go back to that for a second because it doesn't really -- it's not really relevant to whether the measure passes or doesn't.

So, my interest is in trying to make these distinctions meaningful what, if anything are you going to do with the moderate ratings because it doesn't really matter from your perspective when the final vote is tallied, right?

MS. WILBON: Right. It is something

that when in reviewing, in refreshing myself as well on the initial, kind of, rationale for having the high and moderate ratings was so that the moderate rating gives developers a sense of kind of the minimum that needs to be done in order to be considered for endorsement.

And it's supposed to be an achievable rating that can be sought after, that would provide less kind of burden on the developer in order to still be considered for endorsement. So, I think that was the initial rationale.

But I think, Sherrie, you make a good point. I think going forward, perhaps, you know, there will be some -- we could come up with some recommendations and theories from this group about how we should be considering ratings going forward and whether or not there is really a need to have a high and moderate since it's really -- pass or not pass is really the ultimate decision.

MEMBER KAPLAN: Well, I was wondering why -- if you're eventually going to a star rating that makes a lot of sense. But if it's

pass/fail, at least at the Methods Panel, maybe pass/fail is more efficient for us to be considering as opposed to the three tiers. But thank you.

MS. WILBON: Yes. It's a good point. And we don't have any decisions yet. But I think it's something that this group would be really helpful in us thinking through some options and how we might make recommendations to modify that going forward.

So, we'll try to incorporate that into our discussions about criteria recommendations.

MEMBER KAPLAN: Thanks.

MS. WILBON: So, achieving consensus. Again, we're looking for quorums of 66 percent of the subgroup members and we need greater than 60 percent to pass.

Consensus not reached is in the 40 to 50 percent range and less than 40 does not pass. The next few slides are around, kind of, the differences in the testing requirements for the various types of measures.

I know you guys are all familiar with this. So I'm going to kind of skim through this.

But again, if there's something there that you would like me to spend more time on if you have questions please let me know and we can certainly do that. I'm getting a question here in the chat box.

Just bear with me here for a second. Let me make sure that -- there's a question. So, there was a question about the guidance on ratings, whether it applies to only initial endorsement or that initial endorsement does not require empirical validity.

So, yes, that's correct. But if, for example, a developer only submitted face validity for their initial submission, the highest rating they could get for validity would be a moderate.

So, that's how that's interpreted. And hopefully that helps out. If not, please keep asking.

And there's another, Dave is responding to that. Okay, so we're good there.

If there are other questions on the ratings, feel free to speak up and I'll have a brief discussion on that if it's necessary.

Okay. So, again the next few slides here are about, kind of, reminders around the testing requirements based on measure type. So, outcomes, clinical, intermediate clinical outcomes, cost and resource use measures.

For both reliability and validity we only require either data element at the score, measure score level testing. Again, this is our current requirement.

This is actually kind of the core piece of the discussion that we had planned to have tomorrow afternoon as to whether or not we can finalize some of the recommendations on whether or not that requirement should be changed. But again, we are operating under that current requirement.

So, either or, either data element or measure score. And again, face validity is discussed for new measures.

The second element of our discussion for the five tier recommendations is around this second sub-bullet. That is data element validity testing is provided.

We do not require additional reliability testing. If they do this and if the developer does submit a measure with this type of testing we would apply the vote for data element validity to the reliability vote.

And we did have at least one measure this cycle that submitted based on this policy. So, for composite measures we do -- we defined that as both traditional composite and all-or-none measures does not include multi-item scales or survey questionnaires.

We require reliability testing on the composite measure score. And they can also show reliability of the components, but we don't currently require that.

Score level testing is not required until maintenance. And composite measures do have an additional subcriterion. And you'll note

that so for those measures we do ask for revotes of reliability, validity and the composite construct.

And you'll see that reflected in action in one of the measures we're going to review, we did review, sorry. It's not on the agenda today.

For instrument-based measures, reliability and validity testing is required for both levels, both data element and measure score. And I know it's a little bit tricky to wrap our heads around.

But at this point we do allow multiple performance measures under one NQF number. And I don't think we had any issues with this particular bullet here, so I won't spend time on it. It can be a little bit confusing.

So a few additional reminders. Testing should align with the specifications. We do our best to screen for this in the staffing review before we send the measures out to you. We don't catch them all, but we do our best. And

so just another reminder there.

For risk adjustment this will be a discussion for today around the inclusion or not inclusion of certain risk factors particularly clinical or social factors should not be a reason for not passing a measure on validity.

We do -- we would ask you to kind of focus on concerns with the determination and calibration or the methods they used for adjusting the measures which would be a grounds for voting down on the validity criterion.

But we should be very clear about that in our rationale for the voting. If there are comments about the clinical factors and the social factors, that's certainly something that you can discuss and provide recommendations and other comments to the Standing Committee for consideration. But again, it should not be grounds for not passing the measure.

Again, incomplete or ambiguous specifications are grounds for rejecting a measure. And we do our best to offer developers

opportunities to provide clarifications through the response process that we've added in. And so hopefully those are resolved for the most part by this point in the process.

There may be some measures where that still may be a concern and that would be on the table. And careful validity testing is discussed at the time of maintenance for measures that are returning for maintenance of endorsement.

This slide, basically it aligns with one of the background documents that we shared which is the SMP evaluation guidance.

These are some of the recommendations that were shared in that document to provide developers some additional guidance on the things they should be paying attention to when they are submitting their measure which included, you know, more detail in describing the methodology.

Don't just, you know, kind of give words. Provide the formula and any other, you know, information for example that they could to provide you as much context about how they tested

the measure.

Again, more than one overall statistic in reporting signal-to-noise for reliability. So sample sizes, distributions and things like that are always helpful.

And then again more detail in the description of the concept validity and how they selected the measure, the relationships you expect to see, the strength of the association you would expect to see and so forth.

But again, that -- we're not necessarily going to be rejecting measures based on two and three. But that certainly, for number one, if you cannot understand the methodology that was applied to test the measure that that would be something that would be on the table for rejecting the measure if it was inadequate.

Okay, any questions? That was kind of a mouthful, and I just want to provide an opportunity to ask questions. The next section we're going to kind of get into the meat of it here.

I know a lot of kind of fluffy stuff up front. But I am looking forward to hopefully getting into some of the more interesting discussion here coming up. Any questions before we move on?

Okay. So our next section is around some of the challenges that were identified by various Methods Panel members through evaluation of the measures this cycle.

You'll note that many of these have come up not just this cycle but previous cycles and have been the subject of discussion for many of our meetings. Again, we, you know, given the time that we have today, we don't necessarily expect to come to consensus on any of these.

If we do, that would be great and awesome, and we would love to have some level of guidance that we could put forth for future cycles.

The discussion here is generally intended to make sure that as these issues come up with measures we're going to review today that

we're consistent. We're not setting new guidance now that would be applied to the measures we're going to be evaluating this afternoon.

This is merely to make sure that if there are any overarching discussions that we can have now that would help facilitate the evaluation of measures later on this afternoon and tomorrow and make those discussions go faster that we have them at a higher level now and that will help us later on in the meeting.

I also just wanted to point out that one of the strategies we're using today to help move things along is to think about these topics in the context of future White Papers.

We've been talking with the Co-chairs about how we might be able to have a productive discussion given the limited time we have to really dive into these issues. They are very complex, and we just want to make sure there is adequate time to do that.

And we think the best way to do that would really to be able to have a plan going

forward for how we carry on the discussion of these topics in a thoughtful way in a written product.

So, you know, repeating what we did for the Methods Panel paper, as well as the risk adjustment paper and taking these topics and developing White Papers of them. So we will be along the way asking for folks to volunteer if you're interested in participating in the paper.

But also be thinking about what those papers would entail. So what are the key elements of the issue that we would want to address in the paper?

What is some of the guidance that we would want to offer and some of the kind of more strategic challenges that we would want to address in each of these papers.

So keeping that in mind, I did just want to offer Dave N. and Dave C. an opportunity to add on or provide any additional context to that opener if you would like to at this point.

CHAIR CELLA: No. This is Dave C.

You've covered it so well, Ashlie, I'm fine.

CHAIR NERENZ: Dave N. You did, I just want to reinforce the idea.

In these discussions that are coming up now we don't seek or want to go back and change any of the votes or the decisions for measures that are not in front of us today and tomorrow. That's not the goal at all.

The discussions we have now are not going to change the rules by which developers submitted materials on this cycle. There are a couple ones on the agenda later today and tomorrow to which this discussion we're about to have might be relevant.

But a lot of this is really forward-looking and say as this whole process evolves and we continue to think more sharply about some of these issues, how can we evolve these discussions and eventually the requirements for developers and the rules by which we operate, how can they get better and better?

CHAIR CELLA: Larry Glance has his

hand up.

MEMBER GLANCE: Hi. I just want to very, very quickly make a clarification. I just wanted to point out that the white paper that my group headed up was not a risk-adjustment paper.

It was a white paper on evaluating the scientific acceptability of risk-adjusted outcome measures. So it's not just about risk adjustment.

It's about looking at data reliability, date validity, measure reliability and measure validity. It's the whole thing. Thanks.

CHAIR CELLA: Good point.

MS. WILBON: Thanks for clarifying, Larry, appreciate that. Sherrie, do you have a question?

MEMBER KAPLAN: Yes. Just sort of in general terms, I mean I'm not looking for any more meetings.

But it's tempting when we do it this way to discuss these issues first and then

discuss the measures the bleed gets problematic. You know, it's sort of hard to straighten out.

And I'm wondering if it isn't better either to separate the two, this kind of forward-looking guidance for NQF for measures development in the future versus considering current measures in front of us and separate the two things pretty distinctly because I'm finding the bleeds, you know, harder and harder to track.

MS. WILBON: So, Sherrie, let me just make sure -- I'm just going to repeat that to make sure I understand. So you're suggesting that we have a discussion now that's more about how we would -- the question that would help us evaluate the measures coming up and separate that from future guidance?

MEMBER KAPLAN: Well if it's, you know, I mean in a way what we're talking about is forward-looking. It's, you know, what could be done?

What are acceptable methods and measures for establishing, for example, threshold

type scoring versus norm reference tested scores, et cetera, et cetera?

That's the kind of thing that's forward-looking. And then we go into the discussion of the measures.

I was maybe suggesting either we flip the order or at least, you know, it's too hard for me to sort of -- maybe I'm just on West Coast time and I'm cognitively impaired as I get older.

But this kind of thing is tempting to kind of slide into the discussion then we have about existing measures which are going to evaluated based on former guidance not current guidance.

CHAIR NERENZ: Dave N. here. Sherrie, I think that's a good point, and we could probably take the view that much of what we're going to talk about now or the main emphasis is the forward-looking.

Maybe the way to talk about the connection is that a number of the measures this cycle have exemplified or have crystallized some

of these issues. And you're right, we can't treat them differently as a result of discussion we're about to have.

But part of my thinking, a lot of these may have exemplified some of the problems I think we have to try to work through.

MEMBER KAPLAN: Thanks.

MS. WILBON: Thanks, Sherrie. I mean you bring up a good point. And that is our job as NQF staff to make sure that the criteria that we apply for the measures being evaluated does align with our current criteria.

So that's -- I will make sure that we're on our toes for that discussion, but point well taken.

MEMBER KAPLAN: Thanks.

MS. WILBON: So with that we'll dive right in. Just a couple of quick reminders about our current definitions for reliability, data element reliability.

We consider repeatability or reproducibility of the data elements for the same

population in the same time period.

And measure score on reliability is focused on precision and the proportion of variation in the performance score due to systematic differences across measured entities or, more commonly, the signal to noise.

So, Dave N., if it's okay, I will hand it over to you. I'm happy to go over this slide. But if you would like to take the discussion from here.

CHAIR NERENZ: Yeah. Let me just tee it up a little bit, then we'll have max time for discussion. There was an email that went around. I'll just try to summarize that briefly for those of you who didn't get a chance to take a look at it.

This is an old issue for us. We've had this in some of our conference call discussions before we even went into cycles of measure evaluation. It's popped up now and then. But my sense is a lot of the measures that we saw this cycle really call the question on this issue

of thresholds.

You know, if we look back at the 2011 document that serves as the basis for much of what we do so far when talking about reliability and, you know, what criteria we should be looking for, there's reference to vague phrases like so-called accepted standards but without defining what exactly those standards are.

There's a table in the back of that document that does make reference to the .4 is kind of a minimum threshold. It doesn't name it that way. It doesn't come out and declare it that way.

But the number appears. And I think as we watch measures come through and we look at how the developers describe what they're seeing this often comes up.

The reference is usually to the Landis and Koch article from late '70s that was set in the context of the kappa statistic for inter-rater agreement. But others have made different statements about thresholds.

Adams, in the 2009 article for a piece that's often cited about signal-to-noise, talks about .7. Allan Kozlowski did a piece a couple years ago pointing out that there could be even significant misclassification of providers with a reliability statistic of .9.

The Heet (phonetic) article that was given to us this cycle to take a look at makes passing reference to.7. So there are different numbers out there that are offered as thresholds or guidelines without any clear reason, I think in my mind, for saying we should adopt one or the other.

The call the question time comes when you get measures that are bringing forward reliability statistics in the .5, .6 range. And at least in the subgroup I was part of, we had many, many of those this time.

And the action question is clearly what do you do about it? Do they pass or do they fail?

And I noted also in the slide you saw

a little while ago that was defining the ratings for low, moderate, and high, the phrase not satisfactory was part of the definition of a low rating. So what's not satisfactory?

Is something at .5 not satisfactory? So, there's a clear call to question -- this year when you come -- when you see reliability statistics in this range.

Ultimately, do they pass or do they fail? So I think somehow we need to come to grips with this a little more sharply than we have in the past.

And also Sherrie's question a while ago about what's the difference between moderate and high. The definitions don't link those two -- these numeric values.

But in other settings, you know, that kind of verbal label does. We may want to say that if somebody shows us reliability statistics in .9 range we might want to call that high and have that label carry forward into the review process.

We don't currently do that. So I don't have a further clear direction. But I did want to have this on our agenda to invite folks who have been thinking about this to suggest possible ways we might carry this forward.

Not that in the next half hour or so we can agree on some specific number. But questions, for example, like should the same general threshold apply to all measures of reliability or do different statistics lend themselves to different thresholds?

Should we ask for different higher or lower standards for measure score than for data element? Are there certain criteria for data element reliability that set an upper limit on what we can expect to see at the measure score level?

So I think these are in front of us, and I'm hoping the people who have been thinking about this a bit.

CHAIR CELLA: This is Dave C. I'm going to call on Jack because he's got his hand

raised.

MEMBER NEEDLEMAN: Thank you. That was a great summary. And I just want to add two or three things to it.

One of the other issues that has emerged in some of the reviews is our developers are picking threshold levels. And, by the way, the documentation we've been getting has been extremely good.

Nothing I say about the measures is a slight at any of the developers. They've been doing a great job of trying to document things. And as we've asked for more, they've been incredibly responsive. So the developers have been terrific. It's about the measures.

But one of the other issues that I think has emerged here is the thresholds for the number of cases or the size and practices where we're getting reliability statistics across different ranges of those numbers.

And the overall reliability may meet an acceptable standard. But for the smallest of

the groups they don't look like they do.

And the question is how to deal with reliability when we think the measure developers picked a threshold that's simply too low in terms of when they start, who they are including within the measure. So that's another issue to throw on the table.

With respect to the Landis-Adams documentation here, I would simply note that Adams when he said .7, analyzed things at .7 and found that in a simple split of upper quartile rest of the sample, .7 produced -- was a number in which 20, 25 percent of them, if I'm remembering the article correctly, could be shifted from one side of that dividing line to the other which suggests that -- and we've got to think about how comfortable we are saying a measure is reliable when 25 percent can shift.

Similarly, the Landis article these classifications were arbitrary in that. They did not ground their definition of these moderate, substantial, almost perfect to any empirical

analysis.

They had simply picked those labels based upon their -- some -- without analyzing the impact on reliability or so forth. So I think there is a real issue about whether the Landis characterization of these levels should be accepted by us.

And I personally think that the Landis standard is too low.

CHAIR CELLA: Thanks. David, just as a process, there's an interesting quirk on the hand raise function where as the list comes up they are ordered by alphabetical order by first name.

So Alex and others are going to be advantaged if we don't watch out for that, and Zhenqiu is never going to get to say anything ever at all. So I'm going to try to watch the sequence as best we can.

I think actually in this case I've got Alex and then Mike Stoto and then Larry. But I'll try to watch the sequence in which they pop

up.

CHAIR NERENZ: That's correct. At least we have correlation of one on that.

MEMBER SOX-HARRIS: Great, so this is Alex --

CHAIR NERENZ: And then Sherrie is next after that. This is Dave. Go ahead, Alex.

MEMBER SOX-HARRIS: Okay. I really appreciate the opportunity to discuss this. I think this is obviously an essentially important issue for this group.

Unless we have some consensus or, you know, reasonable agreement on this, it's very difficult to make judgments on the measures. So I think this is great that we're doing this.

So I went back to the Landis article to really understand what it said and did and what it didn't. And it was clear to me that it's a really -- it's addressing a very particular context.

It's element reliability for sure and then it's a particular flavor of that which is

agreement between raters on a categorical classifier. And it has its own range and distribution of data.

The statistic kappa has its own range and distribution including negative, you know, numbers below zero. It's testing a specific hypothesis of -- null hypothesis of no agreement.

And as Jack just said, the proposed mapping of kappa to adjectives was completely arbitrary and not evaluated in any serious way.

So even if it was the last word on this particular context, you know, the agreement between raters on a categorical classifier, it's a conceptual and mathematical leap to other reliability contexts.

And in fact, just in my little literature review of previous thinking and work on, or more recent thinking and work on this particular context, it's not the last word. There are people who have evaluated those ranges and found it to be problematic if you apply those ranges. And after my comment I'll put a link to

one such paper in the chat box.

But the overall point I'm trying to make is that I think we need to think more carefully about reliability context, not just the element versus measure distinction, which I think is essential.

But even within, like element reliability there are different contexts. There are different data generation methods. There are different statistics that I think we need to be more specific about.

And just as an example of another element reliability context, think about reliability of an instrument. Like I know some of the measures that -- well one of the measures we're looking at today has the KOOS and HOOS instrument to measure hip and knee pain and functioning.

So, and I believe the reliability statistics was something like a test, retest reliability of that instrument. So I mean there are dozens if not hundreds of articles to give

guidance for what are acceptable ranges of test -- test reliability of an instrument like that.

And none of them even overlap practically with the Landis scale. You know, .7 is considered minimum commonly.

So I think in that context we would be charged looking forward to look at instrument level reliability and what is the current best thinking of that separate from the agreement between rater context and so forth.

And finally, and I apologize this is a little long-winded. But, you know, and then jumping into a completely different context which is the measure level reliability where the distribution of data and statistics are again different.

And so we need to look, I believe, to work that is specifically evaluated, what different reliability levels mean in terms of misclassification and so forth in that specific context.

Not just, you know, take one mapping

of statistics, adjectives in a completely different context and assume it applies across all contexts. So I'll stop. Thank you.

CHAIR CELLA: Great. Thanks, Alex. I think we have Michael next.

MEMBER STOTO: Hi, yes. Thanks. I think those are both good comments. And I think mine relates a bit to what Jack was speaking about. And the idea is that reliability really depends, in practice, on the number of observations that go into the measure. You know, if you're looking at what physicians did certain procedures, you have to look at how many encounters. If you're looking at a consumer satisfaction survey, it depends on how many -- what was the sample size for that one.

So, you know, we tend to look at this in terms of testing, whether there were sufficient numbers in the reliability tests that are presented to us. But that doesn't really answer the question. If we say something has good reliability in general, or in average, the

question I think we need to address is, when it's actually used for a performance measurement, were there enough elements, however defined, to have a reliable score?

You know, I think what we need to say is, you know, this only works if you have so many encounters or a sample size of x and so on. And really to set thresholds for when these measures are used.

CHAIR CELLA: Okay, thanks. That's good. I think we've got Larry, then.

MEMBER GLANCE: Hi. Great discussion. I've got a couple comments to add to this.

I completely agree that that Landis scale is completely ad hoc. Also suspect that although the Adams criteria greater than .7 is slightly more based on empiric analysis, I suspect that if you were to look at different measures and different parts of measures that you would see different thresholds for the signal-to-noise ratio to be when you see physicians switching from one group to another in terms of

performance. In our white paper, I think that we kind of suggested a level greater than .7 for reliability which was reasonable. And I think that may be a good starting point for this discussion.

The second comment I have is that when you're looking at data reliability I think that it's important to consider which data element you're looking at. I think it's critically important that the outcome data element be highly reliable, because that's the essence of the measure. If the covariates that are used the risk adjusted model are less reliable that's not as much of a problem.

So, when we're talking about data reliability we may want to think about having different thresholds for the outcome versus standard risk factors.

And the third and last comment that I want to make is that -- and we talked about this quite extensively in that white paper -- is that a risk-adjusted measure that is poorly risk-

adjusted, meaning that you're not accounting for differences in case mix between providers, hospitals, and physicians, you may appear to have pretty good reliability. But then when you start to really do a good job of adjusting the differences in base mix, all of a sudden you will see that the reliability will go down substantially. Meaning that you can't really evaluate score level reliability independently from score level validity. If you do a poor job with risk adjustment it will end up biasing -- it will upwardly bias your estimate of the measure reliability. Thank you.

CHAIR CELLA: Thanks, Larry. I think we're moving along here. I've got, at the moment, Sherrie, Jennifer, and Eric.

MEMBER KAPLAN: Hi. I want to disagree a little bit with Larry, unless I misconstrued your point there, Larry. But you can be reliably wrong. And so the idea that, like my bathroom scale is reliably wrong, this is the kind of thing that we need to kind of see in

context.

And so, the idea that reliability can be uniquely and kind of we can do thresholds sort of without paying attention to the type of measure we're using and the purpose it's being put to is I think a problem, at least for me.

There's all different kinds of reliability. Inter-rater reliability, test-retest reliability, internal consistency reliability, split half reliability. It depends on what you're trying to measure and why you're trying to measure it, the purpose you're trying to put it to.

And I think that's important for things like between physician differences. And we sent around a while back some work that we've done showing the variation for different types of measures, like attributing to the doctors the number of office visits made, for example, for efficiency purposes.

And the variation around that is very, very big. You can't discriminate between

physicians using that kind of measure. The error variances are too large. On the other hand, for composite measures, you can narrow the variance very tight and then risk adjust those and then you can narrow the variance so that you could use those kinds of measures.

So, I don't think that we can establish sort of what's thresholds uniformly across all different types of measures for all different types of purposes.

But, having said that, I would also say that the units of analysis are a problem here for us. And we need to be thinking carefully about them, especially for things like composites, like functional status that we're going to review, because you have another term in the denominator for certain kinds of units of analysis at the physician level, for example. You have within patient across items in the denominator. Then you have across patients within doctor. And then you have between doctor variance all in the denominator.

That reliability coefficient, which in this case would be an interclass correlation coefficient, is going to be low. It's not going to meet the targets even probably here. I suspect that it's going to be down around .05.

And so I think that those kinds of calls that we need to make about what's an acceptable threshold are not going to be uniform. It's going to depend on the type of measure being considered and then whether or not you have data element reliability, for example. And I still get a little confused about that. But if you've got disagreement between two raters that's very low, that rating is low, and if you square those Landis coefficients you get the amount of reliable variance. And you can see that. So .4 to .59 is about 16 percent. It's roughly 20 -- so 36 percent. So, that's considered moderate.

But, you know, it depends on the purpose you're putting it to. But if it's at that level and then you roll it up to the next level of the score level and you've got

imprecision there, you've got a real problem when you're kind of combining that and then accelerating.

And I put it to David Nerenz's piling on the reliability problem for the error variance. We've got some issues around units of analysis that I think are important to consider.

CHAIR CELLA: Okay, just David here. So far, Sherrie, even though you mentioned perhaps a little bit of disagreement with Larry, I'm hearing largely so far some common themes about thresholds or expectations being context-dependent, statistics-dependent. You can't say that a guideline or threshold for one automatically then carries to the others. And that's with respect to what we said in the white paper that Larry led about, you know, .7 is at least a rough general threshold of expectation. So, I think we can work through that in a bit more detail. But at least so far I'm not hearing sharp and clear disagreement.

We need to start out -- so I guess so

far on the list we've got Jennifer, Eric, and Sam.

MEMBER KAPLAN: Dave, can I just kick back in for one quick second?

CHAIR CELLA: Sure.

MEMBER KAPLAN: Sorry. The Adams thing, the .7 came from the group comparisons from Nunnally (phonetic) originally. And so it was thought, well, if you can't at least achieve a sort of 50 percent reliable variance you're not really doing very well. And it came from Nunnally a long time ago. And I think that needs to be reconsidered too.

CHAIR CELLA: Okay, thank you.

MEMBER PERLOFF: This is Jen. Just a quick comment to repeat something that was said earlier about the minimum sample size at the physician and provider level.

One of the things we see in resource use and cost measures is that developers will choose a bottom threshold, often 20 or 50 events or episodes at the physician level. And they

obviously make a trade-off. If they lower the threshold, the more physicians, clinicians can be included in the measure but the lower the reliability sometimes.

So there is this tension. And it would be helpful to see developers test different lower bounds on the minimum thresholds for participating in the measure. So I just want to throw that idea into the mix.

CHAIR CELLA: Okay, thanks. That's good. Eric.

MEMBER WEINHANDL: Yeah, thank you. So, I was in Subgroup 3 this time. And one of the interesting facets that I see in several of the measures, and I think they were common to a single developer, was an appeal to a quantity profile IUR. And it was new to me, so I had to read the literature about what that quantity was. But, if I understand it correctly, it was essentially about being able to identify reliably outliers, extreme values in a distribution of measure scores.

And so I sat there and thought to myself, well, how do I evaluate this in the context of the guidance of the SMP? Because here we're talking about, you know, essentially reliability statistics for individual measures that look to be pretty modest or poor on some of these scales that we're talking about. The profile IUR appears to be quite good.

And so then it raises the question of, well, if we're talking about reliability, do we need to be thinking about reliability in the context of the intended use of the measure or what, say, a payer might be interested in grading facilities, physicians, et cetera, on with this measure?

That seems to me to be a little bit of a tenuous strategy because you can, of course, talk about how the measure is going to be used today as proposed. But what if the measure -- which, you know, every facility or every physician -- is using for an entirely different purpose in a future application such that, you

know, worrying about outlying values becomes less of a concern?

And so then, you know, the interest should rotate back to, say, the inter-unit reliability rather than a so-called profile inter-unit reliability.

So, I think that, you know, the general point that I'm trying to make is essentially that, in the context of all this discussion about what are appropriate threshold values for reliability, it's the question of, well, if we have a measure, how will it be used? And does that influence what kind of reliability statistics ought to be considered?

CHAIR CELLA: Eric, thank you. If everybody could just sort of hold that thought in mind. And I think that is really the essence of our next block of discussion that appears on the agenda. And I also was part of the same subgroup and had the same sort of observation and question. So, rather than follow immediately if we can just hold that we can delve deeply into

that when we get on that guideline.

Sam, I think, is next.

MEMBER SIMON: Hi, thanks. Yeah, just very quickly I wanted to go back to a point that I think was originally raised by Alex around context. And when we are thinking about data element reliability I think one thing to keep in mind is that the prevalence of the data element is going to matter quite a bit in terms of the maximum amount of agreement we might see.

We don't often see this contextual information, but I do think it would be helpful to us as evaluators. So I just wanted to sort of put a pin in that and raise it to the group. Thanks.

CHAIR CELLA: Okay, thanks. I've got Zhenqiu and then it looks like we have Sherrie again.

MEMBER LIU: Yeah, hi. This is Zhenqiu. So, obviously, we are all concerned about low reliability, and that's understandable. But I also want to draw attention to that

sometimes like in a monthly cycle in the past I have seen in several measures reported better than one. Several measures and when they calculate reliability for that measure it's about 25 percent or even 50 percent with measure reliability as one. I mean, just I'm very skeptical.

CHAIR CELLA: Let me just follow up on that. I can think of a couple where we've seen that. Is there any way that we could possibly raise questions about a "too good to be true" sort of situation? How do we think about those when they come in front of us?

MEMBER LIU: I think it depends on content also. You know, I would prefer more detailed information, more nuanced information. So it may just show you something that I calculate this in that certain method. And certain methods tend to give you higher number. You know, I have seen the same data using different methods to calculate you can get really different. And, you know, some can get you .8 or

.9 and you use a different method you get .3, .4. So, it's not easy -- I'm not sure there's an easy solution.

CHAIR CELLA: Okay. So, Zhenqiu, thank you. I've got Sherrie and Alex.

MEMBER KAPLAN: Hi, this is Sherrie. The issue with the sort of really, really high levels of reliability seen caused me to wonder, and this happened a few times a few cycles ago too, where it looked like the measure developer was averaging at the patient level across patients just averaging the patient level data for each hospital, which isn't how you, you know, would do hospital level, for example, interclass correlation coefficient.

So, I think that's come up before. And it does raise questions about it. But it goes back to detail. If we could get more detail about exactly how things were being done, what measures were used, and how they were being calculated, that would help us to understand whether or not we can believe what we see.

CHAIR CELLA: Okay, Alex.

MEMBER SOX-HARRIS: Yeah, to continue that thread, I mean the whole -- in research it's more common for journals and others to ask for source data and statistical code to enhance transparency and evaluation of what people actually did. It's a whole open science framework that's similarly just asking for more access to the details.

So, that's something that could be considered. Is that some kind of minimum data set? And actual statistical code, the computer code that was used to generate the statistics, so we can answer these questions more directly about did they do it right.

CHAIR CELLA: Okay. Jeff.

MEMBER GEPPERT: Yes. Just a question for the methods folks. So, does it matter if the measure uses some kind of shrinkage to the mean? Does that change the way that we think about the reliability metric or any potential evaluation of the reliability metric?

CHAIR CELLA: Sherrie, do you have an answer?

MEMBER KAPLAN: Well, any time you inflate the mean you're obviously going to reduce the variance. But you probably spuriously inflate the reliability as well because you're inflating the mean and shrinking the variation around the mean, which is going to usually help you with precision.

CHAIR CELLA: Jack.

MEMBER NEEDLEMAN: Yeah. What the shrinkage methods do is basically they say we acknowledge that the folks with smaller sample sizes are inherently less reliable. And they move them towards the mean. They average the mean of the overall sample with the value for that individual unit. And the rating of that averaging depends upon the sample size.

So, that's what's going on with shrinkage, basically, which means that the smaller practices are all pulled in closer to the middle. And, you know, I sat on another panel

where one of the folks complained bitterly about losing information about individual units. But the fact is that we've got low reliability in those small units. We see that when we get these tables by size, the quartile by size, and we see the reliability statistics much smaller.

So, they're trying to deal with the problem of reliability smaller in these units. The question is, from our perspectives endorsing, whether the shrinkage method accommodates our concerns about low reliability for smaller n's within practices or whether we'd like to see the thresholds simply increased, and whether the shrinkage is an inherent part of the measure or something that's done later on when people are applying it in practice.

I can't remember seeing explicit statements in any of the measures I've reviewed that said "we're going to be using Bayesian shrinkage here."

So, I think those are the issues that are raised for us in looking at that. Is that an

acceptable way to deal with the inherently low reliability of low volume units? Or do we want to see the thresholds raised? And do we need to see an explicit discussion of the impact of that for the reliability?

CHAIR CELLA: Okay, thanks. Paul had his hand up.

MEMBER KURLANSKY: Can you guys hear me now?

CHAIR CELLA: Yes.

MEMBER KURLANSKY: Okay. So, this issue of shrinkage is something that comes up in the Society of Thoracic Surgeons database. And it's something we struggle with in terms of the reliabilities of the measures for small programs.

And one thing that might be helpful is if there was some way, in such situations, for the developer to show reliability over time. In other words, you know, just in a simple explanatory, you've got a small program that had a zero percent mortality on a given year, so the shrinkage would not assign that zero percent

mortality. But if you were actually to follow that same site for three or four years and they had zero percent mortality you would tend to believe it.

So, I'm just wondering, in situations where shrinkage is used, if there could be some sort of additional requirement to show for the small volume sites, for whatever is being measured there, the reliability over time of the measure.

CHAIR CELLA: Okay, good point. Patrick has a hand up. Is Patrick on mute?

I guess not. David here again.

I'll quickly pass this question to our NQF staff folks, and recognizing our time block is about to close on this topic.

A lot of the discussion that we've had, and I have a full page of notes here so far, might be cast in the question of, what's the minimum requirement? Essentially, what's the pass/fail cut-off?

But if we go back to our ratings we

still have this moderate, high. In the future, as you envision it, is there any possibility of our group coming up with some kind of numeric criteria for assigning values of moderate versus high? And if we did that, would that make any difference downstream in the other steps of the process, or even all the way out the back side to the way measures are identified as being endorsed?

Is it conceivable that trying to work on some kind of distinction like that would matter in a tangible and positive way, or should we focus strictly on the issue pass/fail?

MS. WILBON: Hi, Dave. This is Ashlie. It's a good question. I don't know that I have an answer. I would say to the -- I think my first reaction would be: simpler is better. Yeah, without thinking about it a little bit more, I don't know if I have an answer right now.

But I think both are potentially viable. So, I think, you know, again my gut reaction is simpler is better, that maybe a

pass/fail might be more digestible, for not only the Panel but for external audiences. And it might be a good place to start and then if there's more details we can filter that.

CHAIR CELLA: Okay. Seeing that Patrick is having trouble with sound let's do Daniel and see if we can get Patrick back on before we close out.

MEMBER DEUTSCHER: Yes, thanks. This is Daniel. I just wanted to note that in some cases the extremely high reliability could also be related to very high sample sizes analyzed.

And I think we're seeing some of these examples through our last cycle. So, for example, analysis as a health plan level.

So, I just wanted to note that's probably also one possibility for those extremely high reliability results.

CHAIR CELLA: Okay, thanks. Larry.

MEMBER GLANCE: I just put my hand down, thanks.

CHAIR CELLA: Okay. So, I hate to cut

Patrick off here. Any chance Patrick is reconnected?

MS. WILBON: We're troubleshooting now. We're going to try to work in the background on that. But I don't think he is on yet.

CHAIR CELLA: Okay, because otherwise just watching the time and making sure we stay tight, I've got quite a good set of notes here and I appreciate the discussion and the thought that you've put into this.

What I can commit to doing is to summarize the notes and try to organize it in terms of a set of issues that we might look more deeply into and at least potential, some steps forward because clearly what seems to be happening here is that we're moving away from anything like a blanket acceptance of .4 as a minimum standard.

I haven't heard any support for that at all, which is a pretty significant thing actually. And I'll summarize that.

And we can then sort of offline by email take up some next steps of, you know, who is specially knowledgeable about this, who is interested, who might want to work further on it. But I appreciate the thoughts and this has been useful. Thank you.

We'll try to get Patrick on and we'll get connected sooner or later and, Patrick, I won't forget you.

MS. WILBON: Thanks, Dave. And, Patrick, if you are able to type your questions in the chat box we can at least try to read your question out and maybe or comment and at least have an idea of what you would like to say.

And we can come back to it. And that's another way to communicate if you're not able to get your audio. So, we'll keep moving.

And as Dave mentioned earlier, the next set of discussions was around, for reliability was around the IUR and PIUR which was something which was a reliability statistic we saw in several of the measures.

I believe that were in Subgroup 3. And I think there were a lot of, we got several comments not only in the preliminary analyses notes but also from the Panel Members who emailed us directly about wanting to have this discussion about how we should be interpreting this and particularly how the PIUR should be evaluated.

There were some measures that provided the PIUR and some that didn't and how that should be, how that should be interpreted.

The other question I think that was posed in some of the analyses and from some of the Methods Panel Members after their reviews were whether or not this particular test or other types of tests only demonstrate reliably for a particular purpose.

So, for example, detecting extreme outliers or demonstrating stability and whether or not this is a method that would be acceptable for demonstrating reliability of a measure that may be used in any context for any purpose given that our current criteria and process does not

provide endorsement for a particular use, but just around the use for a particular, for any accountability purpose.

And then I think the final issue which we've already touched on quite a bit is around the idea of volume and minimal sample sizes and reliability testing and whether or not the testing results should be explicitly associated with the volume or sample size.

So, I'll pause there and I'll hand it back over to the Daves to continue this discussion. I did just want to point out that one of the background materials that we shared, the C.F. Howe (phonetic) article talked in detail about the PIUR.

Hopefully you found that a helpful reference for this question. And, Dave C. and Dave N., I will hand it back over to you guys to get this discussion going. Thank you.

CHAIR NERENZ: Dave C., this is Dave N. I think we didn't choreograph this in advance. How about if I continue this one and

then you get the next two? Does that work?

CHAIR CELLA: It does. I think this is all yours. Go ahead.

CHAIR NERENZ: Well, now that's not quite true. No, I think this is -- and Eric come in earlier about this a couple times hinted at this.

The, a number of measures this time had this two-part analysis where they had IUR and PIUR almost in every case. The IUR was not impressive. The PIUR was pretty good.

And the conclusion would seem to be then the measure is not very good for doing things like putting people in quintiles or giving star ratings or comparing Provider A to B, somewhere in the middle of distribution.

The measure would be reliable enough though to identify extreme outliers. That's not the kind of rating or the kind of decision we're asked to make.

The practical question is what do you do with this? You know, we can take that

combination of statistics and say the measure passes and it leaves our hand saying it's reliable.

But clearly everything we've looked at says it's reliable for one thing but it's not very reliable for another thing. So, it's kind of an open question here, what do we do with this?

And I think, you know, NQF staff and experts are certainly welcome to weigh in because the -- I think we are only allowed as a practical thing in what we've done this cycle and what we have yet to do in the next day or so, we have to use the criteria we're given.

We have to use the categories we're given. But is that going to work going forward in a situation like this?

And how can we handle something like the ones we have in front of us, good enough for one thing, not good enough for another purpose? I have Alex, Eric, Michael, Paul.

MEMBER SOX-HARRIS: This is Alex. One

conceptual shift that might help us instead of thinking of this as two uses of the same measure is to say these are two different measures, related measures but they're different measures and need to be evaluated separately.

And if we think of measures as, you know, specified with an output, one output is just a distribution of, you know, performance. And the other is okay, a determination of whether each individual site is an outlier or not.

So, I think if we can put it in separate measures we can evaluate them separately. The analysis, validity would need to map onto the actual structure of each measure and that might solve the problem.

CHAIR NERENZ: Eric.

MEMBER WEINHANDL: Thank you. And sorry for jumping the gun on this topic earlier. I wasn't looking ahead on the agenda to realize the level of detail.

And so, I like what was just proposed. I wanted to ground it in a little bit of reality

insofar as I understood some of the measures that I typically encounter in the analysis phase.

I think that the challenge is that we truly don't have any control for how the measure is used. And in fact, the measures -- The individual measures are used for both profiling the entire space of say thousands of facilities being evaluated and in other applications they're used specifically for outlier detection.

So, some of the measures it's not as though we are in front of the SMP in this cycle. So, your standardized mortality ratios, hospitalization ratios for the healthcare facility.

On the Compare website somebody could use this when a consumer accesses the website and they see evaluation of facilities that are within their region or their area or their zip code they'll simply see text labels that correspond to as expected, greater than expected, lower than expected.

So, essentially there is text labels

that correspond to outlier detection. So, that said, all right. We'll profile IUR. That will be a great new thing to track if that's what you're using the measure for.

On the other hand, those very same measures like standardized mortality and hospitalization ratios are used in star ratings and they're used in other kinds of payment system that Medicare operates under Parts A and B.

And in those systems it's the point estimate. It's literally only the point estimate.

There is no consideration of the competence or the precision of the point estimate that's used when facilities are slotted in through distribution and assigned a certain point value that rolls up into a summary support measure.

So, I guess I'm showing my cards here. But I think that the practical reality is that we have no control. And in fact, these measures truly are used for, entirely just for purposes by

in this case the same payer, Medicare.

So, I think it's -- It would be disingenuous for us to say well, as long as one or two of the measures or one or two of the reliability metrics are accessible, then we can proceed with providing cover for the measure.

CHAIR NERENZ: Okay, thanks. Mike.

MEMBER STOTO: Yes. I think this is going to be a really challenging discussion. I personally think we actually have to challenge the assumption that NQF makes that we either endorse a measure or we don't.

I think that what we've just been saying, you know, of how it depends on how it will be used and how it depends on the sample size essentially, means that it really isn't scientifically justifiable to say this is either valid or this is either endorsable or not.

And so, I think that the alternative is to say this is endorsed for certain purposes or this is endorsed with minimum sample sizes or things like that. Basically, to work them into

the specifications.

I recognize that might be a big change. But I think that we need to consider it.

CHAIR NERENZ: Yes, thanks. Ashlie and others, any quick response to that? I know this isn't the first time this has come up.

But for the next two minutes as we talk about this should we spend time thinking about how that might be done or is that forever off limits?

MS. WILBON: It's a tricky question. I will say that NQF has certainly been through iterations of thinking through endorsement for particular purposes.

And it's, you know, something that we don't currently do. And so, we haven't actually ever landed on that as a strong path forward because there are just so many considerations and nuances.

And we haven't ever, I think been able to reach consensus at the level of stakeholders and various leadership bodies to go in that

direction.

I think, I realize it does present a dilemma of kind of the reality versus us trying to, you know, have an endorsement type period that is agnostic to use.

But I think at this point I think that should be our path forward because we don't -- we wouldn't be able to provide any guidance at this point on kind of the concept of this infrastructure in which, you know, recommendations around, you know, reliability testing or any other types of recommendations for measures around a particular use of it.

So, it is a tough question. Again, I don't have a specific answer.

But I think the safest place is to be where our current policy lies which is, you know, testing, you know, should be submitted and presented in a way that there is confidence from the reviewers that if the measure is used in any accountability purpose that it is reliable for that, you know, for that use in the context of

any particular accountability purpose.

So, yes. Apologies I don't have anything more specific. And I think that's just probably a question of where we are in this space at this point for endorsement.

Yes, I think that's it. I'm not sure that was very helpful.

CHAIR NERENZ: I know it's a tough issue because we're talking about a fairly significant change in NQF policy and past practice.

It's just, I think it's in front of us that a number of people wanted this discussed because the developers themselves essentially in their presentation told us reliable enough for some things but not for other things.

They're sort of compelling us to think it through. I see Patrick with a hand up. Hopefully we've got the sound issues worked out now.

MEMBER ROMANO: Yes, can you hear me now?

CHAIR NERENZ: Yes, we're good. Thank you.

MEMBER ROMANO: Okay, great. Yes, sorry. I've just been waking up on the West Coast and I thought I was on Zoom. So, I was not dialing in properly.

So, anyway I wanted to make a point I think others have here that as we think about inter-unit reliability really we shouldn't be thinking about a threshold.

We should be thinking about how the reliability measures provide guidance in how the measures should be used. It's not a question of whether the measure is acceptable or not.

It's a question of how to use the measure in a way that it becomes acceptable. And so, we've already talked about some of these suggestions.

So, I could easily as a measure developer manipulate the inter-unit reliability or I shouldn't say manipulate, maybe enhance is a better term by, for example, using two years of

data instead of one year of data to do the calculation or by setting a minimum volume threshold and saying that this measure should only be used for providers of greater than a certain volume or as Jeff suggested earlier, you raised a very interesting question.

Didn't really answer it. But the point is that once you do shrinkage the classic signal-to-noise estimation approach doesn't work anymore because you've used that approach to actually shrink or reliability adjust the measure.

And so, that is a way of addressing the problem. But we have to realize that we are making a tradeoff then. We're making the measure more reliable by shrinking it toward the mean for small units.

But in so doing we're compromising validity and there as we're introducing bias. And so, we accept that tradeoff in certain applications.

But we might not want to accept that

tradeoff in other applications. So, I think that as we understand this problem of inter-unit reliability better it really does challenge the traditional assumption of, you know, whether the NQF endorsement is a stop-go, red light/green light kind of process or whether it's really an issue of reliability particularly about providing guidance to users about how to use the measure in a way that it has acceptable reliability in practice.

CHAIR NERENZ: That's good, thanks. I currently have Paul, Larry, Sherrie in that order.

MEMBER KURLANSKY: So, I think this is extremely challenging issue which I struggle with philosophically and practically. And it came up in this context with the PIUR.

It also came up in context with the socioeconomic risk adjustments, et cetera. And it's extremely difficult.

I mean even if we were to provide advice to how a measure could be used,

unfortunately I think, the more I think about this I think there is probably no ability to actually control what will actually be done once the measure has received some level of approval, which makes me tend to think that we should apply sort of a minimum standard, if you will or maybe that's not the right term.

But in other words, the measure should be considered, the threshold should be high, in other words.

There, for every application that we can see this reasonably being used what would be the reliability rather than the low threshold which would be well, for this particular situation, you know, to identify outliers it's okay, but to develop and distinguish people who are in the vast majority of the middle it is not reliable.

Therefore, we would approve it for this particular purpose. I don't think we can practically do that.

And so, I'm tending to think now that

perhaps we should be very, have a very high threshold and say or any, you know, application for which this could reasonably be applied it would have to be, they would have to demonstrate, the results would have to demonstrate reliability.

MS. WILBON: Hi, Dave. This is Ashlie. If I could just jump in quickly. Jack and I were actually having the chat on the webinar about this exact issue.

And his question was, you know, am I implying that given that we don't endorse the specific uses of reliability just based on the potential use that requires the highest reliability.

And I think what Paul just stated is exactly kind of what I would say but he stated it much more eloquently. You know, given the spectrum of accountability applications and we don't know where the measure might be used and as it's been stated we don't know how a measure might be used by any program or we can't predict

that in the future that, you know, we should thinking about evaluating the measure that is good enough for any accountability application.

And that may be the highest level of accountability down to the lowest level of accountability and that we should kind of be thinking in that context at this point in time.

And hopefully that's helpful. And I think again, what Paul stated I think is very much in alignment with what we would like to be thinking here.

And so, based on where we are at this point not having endorsements for particular purposes. I think, you know, the other challenge is that, you know, tagging measures before, you know, particular purposes often gets lost.

So, when measures are endorsed, you know, there's an asterisk next to it or, you know, special consideration that this measure should only be used for x purpose. That gets dropped, it gets lost.

We had in the past a time-limited

endorsement where measures that didn't have testing yet could have a time-limited endorsement. And so, what we found is that what people kind of latched on to was the endorsement piece.

And all the other kind of asterisks and details about the conditions of that endorsement were lost. And so, that's one reason why we kind of had moved away from that.

But I think having, you know, endorsement means the same thing for all measures is the place that we should, you know, kind of be focusing and trying to think about how our evaluations could support that perception for people who are using an endorsement.

CHAIR NERENZ: Good. I've got Larry and Sherrie and Jack if you want to jump in too after that.

MEMBER GLANCE: I'm a little worried about going right before Sherrie because she's going to respectfully disagree with some things I'm going to say.

CHAIR NERENZ: She can probably --

(Simultaneous speaking.)

MEMBER GLANCE: At the high level validity or reliability or whatever we want to say, so I've got a couple of points that I want to make.

The first one and I think the most important one is I think that we, you know, at the end of the day we need to be kind of pragmatic here.

We're having a really, really difficult time choosing a threshold for just the classic inter-unit reliability or what we often times consider it's signal-to-noise ratio.

And that is going to be tough for us. The profile inter-unit reliability, it's a very interesting way of looking at reliability. This is something that is just now appearing in the literature.

And it uses the abstract. The authors published this about a year ago said we propose an alternative measure of reliability. This is

new.

I don't think we need to establish this as one of our testing criteria for reliability. I think we should let this play out a little bit and see what other people think about this before we adopt this as one of our standard tools to evaluate measure level reliability.

That's my first comment. The second comment I really want to iterate I think this is a lot -- Is that when you look at measure level reliability you first need to convince yourself that the risk adjustment for a risk adjusted outcome measure is balanced because if it's not a lot of the variability between providers, and that's what's going to drive your reliability in part, is maybe because you have inadequate risk adjustment.

So, if you have a, you know, you look at cardiac surgery if you're going to compare quaternary care centers that take care of the sickest of the sick to the community hospitals

that take care of really routine cases and if you don't do any risk adjustment, that's the extreme of broad risk adjustment, there's going to be really good reliability because there's going to be a big spread in outcomes between the providers that have the high mortality rate and the providers that had the lowest mortality rates.

So, I think you need to look at risk adjustment when -- First before you look at measure level reliability.

And then the third point I want to make, I want to push back ever so gently on the comments that Patrick made.

I think that what we call enhancements so the use of volume cut offs, looking at a longer time frame, the use of shrinkage, I think those are perfectly acceptable ways truly pragmatic ways to allow you to measure performance, provider performance in a reliable way.

So, if we know consistently that very low volume providers and those that have case

volumes less than 25 that their performance just jumps around in a very erratic fashion would be what you would expect because of the small sample sizes, it's okay.

And that doesn't mean something is messed up. The second point is same thing with time frame. There is nothing wrong with looking, instead of looking at one year maybe looking at two years or maybe even three.

It's a little problematic when you go out further to the longer time frames because it may be that the quality of a provider changes over time. But that is still something that we have to consider.

And then the third thing is use of shrinkage estimators or reliability adjustment is a very standard approach. And it's being used increasingly by many groups.

It's what CMS does. And there are some heated arguments about using it or not using it. But I don't see any reason that we should penalize measure developers because they use

shrinkage estimators.

That is a very valid and acceptable approach. And what it does is it basically minimizes the fairly, sometimes fairly wide fluctuations that you see in provider performance when the provider volumes are low, okay.

So, those are the main points that I want to make. The first one again, profile inter-unit reliability, I don't think we need to go there.

The second point, consider risk adjustment before you look at measure level reliability. And third, the enhancements -- they're okay. There's nothing wrong with it. Thanks.

CHAIR NERENZ: Okay, thank you. Good place to go. Sherrie, do you want to respectfully disagree or are you going to go a different direction? Is Sherrie on mute?

MEMBER KAPLAN: There I am. Am I okay now?

CHAIR NERENZ: You're good enough.

MEMBER KAPLAN: Okay. So, I wanted to respectfully agree with everything Larry said. And I don't want to agree on the shrinkage estimating rabbit hole.

But I wanted to come back to a point to Ashlie's point. I got confused about what exactly the, you know, purpose of measurement, what you're going to do with that now.

And I'm concerned again about if you're going to use my weight on my bathroom scale for my psychological well-being that's fine. That's a good purpose.

But if you're using it to give me anesthesia, you know, it's wrong and I'm going to get into trouble. So, I really am concerned that NQF, CMS changed the game when they started adjusting compensation.

They adjust, you know, we went from quality improvement when Helen and I were having this discussion years ago about NQF's agnostic position on the purpose of measurement, from quality improvement to quality assessment and

then adjusting compensation.

So, maybe a way forward is to, for NQF is to consider the reliability question for different units of comparison.

While it may be okay at one level of comparison unit like the hospital or something it might not be as good for or reliable for estimating physician performance, for example, comparing -- certainly not comparing individual physicians.

Is that a position NQF could think about or have some basis for kind of adjusting purpose at the unit of comparison level?

MS. WILBON: It could be. I think the challenge is that we're not there yet. And it would take some time to get there.

And again, I don't have a great answer because, you know, it's something that we kind of -- we tried a couple times in the past to think about this issue and how it might impact the weight of the endorsed measures.

It doesn't mean that we can't revisit

it and I think we should. But I just, I get a little bit -- I don't want to have you guys coming up with recommendations or guidance on a future that could be and we're not sure what the context of that is or what the kind of parameters for that would be.

And so, I guess, I think my response would be, yes, it's something we can and should consider. But it would be a future, you know, a potential future state.

MEMBER KAPLAN: Thanks.

CHAIR NERENZ: And actually, Dave N. here. Clearly, we get into highly forbidden territory if we cross the line in starting endorsing or approving uses themselves.

It's not really our territory. Our focus should be on the measures.

But clearly again, examples as with these IUR/PIUR distinctions we're looking at data that tell us that a measure is reliable enough for a certain kind of use, is not reliable enough for another.

It doesn't mean we're judging the use itself. We're just saying, is the measure good enough for this or that.

I'm also, I noticed here as we go along the point Eric made a few minutes ago about, you know, if you look at the portfolios ways in which CMS uses certain measures.

And sometimes in which it's the point estimate and the variance, but other times it's the point estimate only. It's the same measure that's used because it's NQF endorsed.

But if we really pull back the layers of the onion we might say that the endorsement should have supported one or two of these uses, but not have supported the other one.

So, I'm left with the idea that it's hard for us to fully step up with integrity to our task as a scientific methods panel to judge reliability without somehow going a little further and saying reliable enough for x, y, and z, but not reliable enough for p, q, and r.

I understand there are a lot of

details to work through all the way through the CSAC and Board level at NQF about how that might play out.

But I just, you know, I'm sort of going back to the beginning of this that in some of these cases the measures, the cycle of the measure developers themselves have essentially made that statement.

And then we have to somehow deal with that. Daniel has got a hand up. Go ahead, Ashlie, and then Daniel.

MS. WILBON: Yes, a quick question in response to that. Is it, could it be though that this particular task is not one that we would want -- if it is narrow in scope in that way and the idea is that NQF endorsed measures could be used for any accountability application, is it that this just isn't one that we would want to be for measures that are presented?

I'm kind of thinking about it the opposite way instead of us thinking about applying endorsements or a particular purpose

based on the test submitted, but to provide some guidance to say if you were looking for, you know, measures that could be used for, you know, the spectrum of accountability applications that these are the types of tests that could demonstrate that and that maybe this is one of those tests.

Is that something to consider? I'm just trying to kind of digest the discussion as well.

CHAIR NERENZ: Yes, Dave here. I guess just responding for myself, I wouldn't go that way at all because it seems to me that the more information that we get about reliability in different forms and different flavors is just as useful scientific information about the evaluation measures suitable for this or that purpose.

So, I would not ever shut anything off and say we don't want to see it. I think the question is, what do we do with it when we do see it? I've got Daniel and Jack in the queue now.

MEMBER DEUTSCHER: Yes, this is Daniel. What if I ask a question on whether there may be a suitable first step could be that instead of going all the way to an endorsement for a particular purpose which definitely requires more thought and I acknowledge it could be difficult to monitor and control, my question is could NQF request that at least the minimum threshold of number of cases or minimum time period needed to achieve an acceptable level of reliability first identified by the developers and then added to the specification?

MS. WILBON: Daniel, could you repeat the last part of your comment, sorry?

MEMBER DEUTSCHER: Yes, sure. I was wondering whether NQF could request that the minimum threshold of number of cases or the minimum time period or a combination of both needed to achieve an acceptable level of reliability be first identified by developers because we've noted that not all of the developers go ahead and identify those

thresholds?

And then add that to measure specification. It sounds like you would need, you know, a compromise on maybe a first, a suitable first step going forward with some kind of restriction in terms of the use of the measure under certain conditions.

MS. WILBON: Yes, I think that's a good question. It's actually something that we have been trying to do when possible to encourage developers to indicate, you know, sample size requirements and include it in the specifications.

I think that is something very feasible that we could do and it is something that we do try to do. It's not -- and we could probably be more explicit about it in our guidance.

But I think that's certainly a very feasible approach.

MEMBER DEUTSCHER: Okay, thanks.

CHAIR NERENZ: I've got Jack and Mike.

MEMBER NEEDLEMAN: Thank you. We are in a very rich discussion. And flying through a bunch of things on the schedule. I really want to come back to this issue of reliability for use.

But just want to note, shrinkage we ought to have on another call because there are multiple methods there including not ignoring things which, you know, right now most of our shrinkage measures are towards the general mean and the question is whether those methods are appropriate, whether we're losing information about relative performance and we simply average to the general mean.

And we've got risk adjustments later in the discussion. So, I just want to endorse the thought that if the risk adjuster isn't doing the work it's supposed to do, we do not have a valid measure.

But coming back to this issue of reliability for specific uses. I think one of the elements that we've seen in this discussion

is that NQF does not endorse for specific uses.

We've been told that. The question is how that will affect our reliability judgment. What I've seen specifically from new measures is to some extent recognizing they won't immediately be used for payment of the area where people generally feel is the highest reliability is needed.

So, to accept a lower level of reliability for an early endorsement of a relatively new measure to see how it plays out to get more data and so forth.

But given what we've said there is not going to be endorsement for use, the question is whether we should all be moving our standards for reliability to the level that we would require for the use that we think requires the highest level of reliability?

I think one of the consequences of that is a bunch of measures may not get endorsed that would have gotten endorsed in the past.

But we ought to be thinking

individually and collectively about whether standards for reliability ought to be in the most restricted use or the use that requires the highest reliability.

I'm taking from this conversation, that's where I'm going to be moving in my assessments. But we ought to be thinking about that and I'd love to hear some guidance from the other, from the governing boards and to some extent from the developers about their preferences about endorsements for use versus always applying the most restrictive or the highest levels of reliability.

And I think we ought to be inviting people to give us that feedback to help us in how we use the and how we assess the reliability.

CHAIR NERENZ: Thanks, Jack. I've got Mike in the queue. But I've also got Apryl with a hand up. Apryl, do you want to come in on something that should be said or talk about some of the policy issues?

MEMBER STOTO: Well, this is Mike and

I actually do want to do that.

CHAIR NERENZ: Okay, Mike, go ahead.

MEMBER STOTO: Yes. I just wanted to second Daniel's suggestion and maybe extend it a little bit. I think that asking the developers to say what's the minimum sample size I think would go a long way.

And, you know, in a way it's not that different from what we already do because we require that they test it based on data that's related to how it will be used. And sometimes the developers lay out, you know, how the reliability depends on the number of encounters and things like that.

So, I think that we could just try to look at that more carefully and get that to be done more extensively.

CHAIR NERENZ: All right, thanks. Apryl. On mute maybe. Okay, so let's get to Gene, sorry.

MS. WILBON: It's Ashlie. I was going to ask the operator if you, we'll try to get

Apryl on a future line.

CHAIR NERENZ: Okay. Let's go to Gene then.

MEMBER NUCCIO: Real quick, given that we've got about five minutes before we're supposed to have a break. This is a second bullet and I don't know whether we're still on the first bullet.

It falls on the same line. It's not exactly about use.

But when we look at the data that's provided, for example, detecting extreme outliers and discover that the measure is very good at detecting extremes but very poor at detecting differences and nuanced differences between 2 and 9 in our ten point or categories, I think we should always think about the general use of these measures which is for public reporting at the very minimum. And for performance for many of the measures either individually or as a composite.

And so, if they provide -- if the

developer provides different information about how well the measure detects extremes it's very important evidence that it can detect the large majority of differences or differences among the large majority of our providers we should be, I would suggest rejecting that because the normal use for these measures will be for public reporting and/or pay for performance.

CHAIR NERENZ: Good point, thank you. And if I could just paraphrase it may be that if we don't end up ever in the near future specifically endorsing for use we could anticipate the range of uses out there and set reliability thresholds and presumably validity thresholds with endorsed measures for the kind of uses that the developers talk about or we think will happen.

So, it's a matter of where you set the threshold more so than specific yes/no decisions on acceptability for certain uses. Did I go too far in my paraphrase or is that about right?

MEMBER NUCCIO: That's about right.

My point is that as evaluators of the measure we should be thinking about how do they, what's the evidence that the measure performs well in all reasonable situations?

And reasonable situations, can you detect outliers? If they give us good information that they will then fantastic.

But if they can't make a distinction between a 2 and a 9 on a ten point scale then, you know, it doesn't pass basic reliability.

CHAIR NERENZ: Okay. Do we have Apryl connected?

MS. CLARK: Hi, all. Can you hear me?

CHAIR NERENZ: Yes.

MS. CLARK: Hi. So, this is Apryl Clark. I am the acting vice president of quality measurement here at NQF. And so, I have been working with Ashlie on the Methods Panel.

So, it's great to talk with you all today. I just wanted to maybe just chime in about sort of, you know, endorsement for use.

It's actually something that we, the

NQF has actually been thinking about. We have had some other issues that have come up in endorsement committees where we would consider different type of measures.

So, as you guys talked about there are sort accountability measures. There's quality improvement measures. There's appropriate use measures.

And so, many times those measures don't necessarily always fit exactly into our criteria. And our Standing Committees have kind of talked about how you, you know, they think they are important measures but how do they feel?

So, reliability is one of those things that has come up. And so, we've actually been thinking about is there different levels of endorsement or different ways of endorsement?

So, I'm not sure I would call it endorsement for use per se. But I think that's the idea is that maybe endorsement would be for accountability measures and they have a certain like criteria that they have to meet.

But maybe there's a clarification that would be for quality improvement measures because they would have a different level to meet. And then that's kind of how we would kind of think about both the criteria and Part B.

And as Ashlie mentioned, I think, you know, a little bit the challenge in making sure that measures are used in the appropriate way. But it actually is really something that we're thinking about.

So, I think this has been a really fantastic conversation.

CHAIR NERENZ: Thank you. All right, I'm inclined then to turn back to Ashlie. We have hit the bottom of the hour.

Want to respect people's time and biological needs and other pressing things. So, Ashlie is here.

MS. WILBON: Yes, absolutely. Thank you, Dave. And thanks, everyone for really great discussions.

I think we'll be looking forward to

getting hopefully some participation from everyone on getting these ideas down on paper and having more thoughtful discussions about them in future papers as well as future webinars. So, with that again with attention to your other needs, including those biological ones we want to -- We're going to take a break for an hour.

Let's plan to reconvene at 12:30. And so, hopefully you guys will be able to tend to all of your needs.

We're looking forward to coming back and finishing our last few kind of content level discussions before diving into our measure evaluation discussions later in the afternoon, around 3:30 after another break.

So, we'll see you back at 12:30. If you are able to, obviously if you want to hang up. You might want to stay on the webinar. We'll keep the webinar open so that you don't have to go back that process.

But keep in mind that you will need to dial back in and to give yourself time to get

back on the line so that we can be ready to go at 12:30.

CHAIR NERENZ: Ashlie?

MS. WILBON: Yes.

CHAIR NERENZ: Can we leave the phone lines connected if we don't need to hang up?

MS. WILBON: Yes, you can, you can.

CHAIR CELLA: Ashlie?

MS. WILBON: Yes.

CHAIR CELLA: This is Dave Cella, Dave C. Is it possible for Sam Simon and Dave N. and you to stay on for just three minutes if we could talk about the next session?

MS. WILBON: Also, sure. I also wanted to ask Alex Harris because I realize I inadvertently swapped their names. And I think Alex was the one who emailed me about this issue.

CHAIR CELLA: Okay.

MS. WILBON: But I for some reason put Sam down. Sam, I'm happy to have you still participate and I think it will be great to have a couple other discussors.

But I also wanted to include Alex if you're able to stay on. Apologies for that mix up.

PARTICIPANT: No problem.

MEMBER AUSTIN: So, Sam, you can stay on. Alex, can you stay on and Dave N. and Ashlie? Everyone else can sign off.

(Whereupon, the above-entitled matter went off the record at 11:33 a.m. and resumed at 12:29 p.m.)

MS. WILBON: Let's go ahead and get started.

I wonder if folks from the Methods Panel, if you're on or didn't hang up and you're back on, or you're back on from dialing in, if you could just give us a hello, just so we have a sense of how many folks are on the line?

Or, Operator, could you give us a sense?

CHAIR NERENZ: Or you can raise your hand and lower it, too.

MS. WILBON: Oh, perfect, Dave.

That's a great idea.

CHAIR CELLA: If that was Ashlie, by the way, your voice is very distant.

MS. WILBON: It was. Thank you. Let me try to fix my headset here.

CHAIR CELLA: That was a little better, actually. Maybe just speak up.

I've seen about five or six hands go up. Mike Stoto is sending a message saying he's back.

I think we can get started. People will join in.

And this first topic for the second half of the morning or early afternoon, to a large extent, is a continuation of the previous discussion, particularly that part of the discussion in the previous session that related to reliability demonstrated for a specific purpose. But the specific question came up --

(Simultaneous speaking.)

MS. WILBON: Hi. Could we have folks mute your lines?

Operator, I'm not sure if you can help us with the muting of the lines. That would be great.

CHAIR CELLA: Thank you. Thanks.

MS. WILBON: Could everyone who's speaking please mute their lines?

CHAIR CELLA: Okay. I'll try again.

So, welcome back.

And now, we're going to talk about the relationship between reliability and validity. It may be a shortened session because we covered this topic to some degree in the previous session.

But this was a specific question that came from Alex, actually, although Sam is named on the agenda. And the discussion question was: if the measure has been shown to only reliably be used to characterize outliers, should the validity testing mirror this use? This issue came up, actually, in the previous discussion.

So, Alex, would you like to tee it up specifically with regard to the

reliability/validity question?

MEMBER SOX-HARRIS: Sure. So, this issue came up for me on several of the measures that we were evaluating. And as our previous discussion highlighted, many of the measures the developer presented reliability for either two uses or two forms of the measure, however you want to think about it, one being the inter-unit reliability for the whole distribution of performance scores and the other the reliability for the PIUR for the identification of extreme outliers. And we had a very rich discussion about how we should think about that.

And extending it into validity, and moving on to the validity analyses and evidence that was presented, sometimes the evidence would pertain to the IUR formulation and other times it would be constructed in such a way to map onto the PIUR formulation. So, I think the general question is, how should validity evidence presented map onto the sometimes multiple forms or uses presented in the reliability analysis?

CHAIR CELLA: And, Dave N., did you have a comment you wanted to make, also, as part of kicking it off?

CHAIR NERENZ: No, I think that's enough to get it started. I just was observing that I don't think in the earlier discussion, we specifically linked the reliability and validity discussions, but there may not be a whole lot to add. The question seems to be, if you looked at reliability using a certain approach or a certain kind of purpose, like getting rid of outliers, should the validity testing match? But I think that's already been teed up.

CHAIR CELLA: Good. Okay. So, the floor is open for input, discussion.

It may be that people feel that this has been sufficiently discussed in the previous session. I don't see any hands up.

CHAIR NERENZ: Yeah, I would tend to agree. I mean, if the question is, should the reliability and validity analyses match, I think a reasonable answer is yes.

CHAIR CELLA: Hearing no objection to that reasonable answer, and seeing no hands still up, we may be able to steal a little bit of time for some of the future discussions and move things along.

So, last call for any comments on this issue.

MEMBER ROMANO: This is Patrick.

CHAIR CELLA: Go ahead.

MEMBER ROMANO: One thing I might add is that -- I think this was alluded to earlier -- but in some cases the reliability, the inter-unit or inter-provider reliability could be falsely high because the provider signal is capturing both the variation in risk and variation in outcomes. So, if the risk portion is not dealt with appropriately, then you're essentially exaggerating the provider signal because the provider signal is capturing both features.

So, again, this has some implication, right, because what it means is that these metrics of reliability are to some extent

dependent on our ability to separate confounding. And I don't know how that applies to this particular question, but it may have an application.

We're probably on the cutting edge of research in this field to understand it better, but we know, even, for example, the Adams approach assumes no variation in risk. It assumes that everyone has the same risk. And that's really not a valid assumption for outcome measures. So, the Adams approach is sort of geared more towards process measures, where everybody can do what they're supposed to do. And we have to be thinking a little bit differently for outcome measures.

CHAIR CELLA: That's a great point.

Eric, you've got your hand up. Go ahead, Eric.

MEMBER WEINHANDL: Yes. It's an interesting question. I must admit to speaking of my own way of sort of processing these when I think about hospitals or other kinds of

facilities that are being evaluated.

To the extent that a measure developer draws attention to outlier detection, I certainly am interested -- or maybe I should say that I'm suspicious, of course -- that outliers may actually be more likely to reflect unmeasured risk adjustment or residuals of modeling, just because of unique sociodemographics within a particular area where the facility is located. There may be particular sort of patient characteristics or medical phenotypes that a particular facility is concentrating on, and sometimes with an administrative database those can be very difficult to extract.

So, I must admit that, although I didn't think about it in this cycle as I was looking at these measures that had a profile IUR reported with them, it would be nice to see that, if a measure developer proposes that a measure is focused primarily on outlier detection, that there was some evaluation or some reassurance that these outliers that are being detected also

are as ably addressed by the risk adjustment scheme as sort of overall is.

CHAIR CELLA: A good point again.

Okay. So, I think we should move on. Thank you for that input.

Can we move the slide to the next session?

And we're going to talk now about risk adjustment. I think that we'll focus on social risk adjustment, but probably risk adjustment generally. And, Ashlie, you were going to set this up for us.

CHAIR CELLA: Sure. Thanks, Dave.

Can everyone hear me okay?

CHAIR NERENZ: Still a little faint, but not bad.

MS. WILBON: Okay. Is this better, a little bit better?

CHAIR CELLA: That's better, yes.

MS. WILBON: Okay. Great. I pulled the mic a little bit closer.

So, with this discussion of social

risk adjustment I think we had actually a very, fairly extensive discussion over email just after the evaluations were done. I think that was triggered by an article maybe that Gene shared, or someone. And I think that was an initial start for this discussion, and then, I think certainly after the results for the preliminary analysis were shared back, there was some more discussion amongst the panel via email about the Methods Panel's role in providing evaluation of social risk adjustment.

I think certainly everyone understands at this point that NQF prefers that decisions around the actual factors included in the risk model, including social and clinical factors, are left to the standing committees, although we certainly recognize that SMP members will have substantial and important feedback on those factors as well, given your own backgrounds and expertise in risk adjustment and other areas.

And so, we certainly want to provide the opportunity for the Methods Panel to give

feedback to standing committees on things they should be thinking about when they're looking at the measure in totality, as we look through all of the criteria and in more detail at the clinical aspects of the measures.

And I think the question, then, comes up, so if the Methods Panel is not voting on validity, particularly with an eye on social factors, but certainly has perspectives to share what is the best way to provide that guidance to standing committees and how can we ensure that the standing committees are considering all the factors that were not only recommended in the NQF report, but also ensuring that developers are kind of using methodologies and making decisions on risk adjustments that align with current guidance and research?

So, I'll pause there, and then, certainly hand it over to Dave N. for comments, as he was very involved with NQF's report on social risk adjustment, as one of the chairs. So, I'll just hand it over to you, Dave, for

additional comments.

CHAIR NERENZ: Yes, thanks, Ashlie. I'll just mention a few things specifically about the issue of social risk factors.

As Dave and Ashlie both pointed out, I think the practical questions in front of us apply to risk adjustment in general. The social risk factors are just a specific example. And this was on my mind in the cycle because a couple of the measures that came to the subgroup that I was on coupled social factors in a certain way that raised questions for me. And then, it raised the subsequent question about how do we convey that message. If we think something is not correct, how do we convey that? Where does that message go?

Again, this is not new to us. So, I'll do this very briefly. I think, as most people know, prior to 2015 and 2014, the strict policy at NQF was that social factors like poverty, homelessness, minority status could not be included as adjusters in performance measures.

And then, a panel was convened in 2014. I was co-chair of that. We had a group of experts kind of like the one we have gathered together today, but experts on this issue. We had conference calls. We had meetings. We then issued a report that basically changed the policy or recommended that it be changed.

And a couple of the essential things were that, after we really talked this through, we said social factors should be treated exactly the same as clinical factors in their inclusion or exclusion from risk adjustment models. There's no reason to treat them differently.

Then in both cases you think about whether the variable was present before the start of care. The fundamental constant was that, if a particular variable stands as a marker or somehow reflects differential quality of care, then you don't want to adjust for it because you may be adjusting away quality differences that you want to see reflected in the adjusted measure.

On the other hand, if the variable

reflects other influences apart from quality of care, particularly if the providers being evaluated have no control over it, you do want to adjust because, then, you get the clearest, least biased signal about quality of care.

And so, that report was issued. The CSAC and the Board did change the NQF policy. So now, as you see actually in the text, in the materials that the developers use, they're asked to work through, is there a conceptual relationship for the variable? Is there an empirical relationship, so that there's some association between a factor like poverty and the measure? Usually, outcomes are the ones we're talking about here. And then, ultimately, is or is not included in the adjustment model?

Now what we've seen over the years, since when this report was issued, the public comment was vastly in support of all the national provider organizations' plans, everybody in its court. CMS weighed in against, as did six other organizations.

And so, to this day, we still have situations where a measure is presented to us. All the background work is done, in some cases very thoroughly. Is there a conceptual relationship? Is there an empirical relationship? The answers come up yes, yes, yes. And then, in the end, they say, but we choose not to include this factor like poverty because we choose to.

We never see that for factors like diabetes or depression. If they seem to be associated with the outcome, and it's understood that it's harder to achieve good outcomes with patients with these kinds of problems, they're in the adjustment models, they never get pulled out after. So, there is a differential treatment.

Okay. That's all just background. The reason this is on my mind in this cycle is that in a couple of specific measures I was asked to review, I thought that the analysis of some candidate social factors was wonderful. I mean, it was thorough. It was thoughtful. It brought

in all the relevant data. It did careful analysis. It was great. And all of that, to my mind, said that these factors should be in the adjustment model. Then, the developer said, no, we choose not to include them because we don't want to. Okay.

So now, my quandary, as a reviewer, is, what do I do about that? I'm not allowed to fail the measure on either reliability or validity grounds, except perhaps down Larry's line of thinking, if I say this problem essentially invalidates the reliability statistics. But I'm left with the question, what do I do? I can write a comment, but what happens to it?

So, that's sort of the end of my framing, and I am looking forward to some discussion, including from our NQF experts about what actually can we do and what should we do in situations where we believe the risk adjustment is not appropriate, not good enough? And that would be true if it's a social factor or a

clinical factor.

CHAIR CELLA: Well, that's a really great history and setting up of the issue. And there's a lot to unpack there, Dave, but I think the core question you're asking is, how can we send a message that the standing committee will hear that relates to concerns about not including social risk adjustment when it appears that it should be included? And the only sure way to have that seen would be to vote low on reliability, based on what Larry has pointed out, as you said, that if the risk adjustment isn't right, then the reliability is off. But I don't know if you can tell that from a submission.

So, I open it up for discussion, including NQF, if you want to help guide the parameters of what's possible here, or what's not possible.

Mike's got his hand up. Go ahead, Mike.

MEMBER STOTO: Yes. I think this is a really important issue. I'm not quite sure I

understood the issue about, not risk adjustment being a reliability issue, but it certainly is a validity issue. That's why we talk about a risk adjustment in the validity section. And it seems to me that we should treat risk adjustment as it either makes this valid or not valid, regardless of whether we're talking about clinical issues or socioeconomic factors.

If we think, and the evidence suggests, that there really are sociodemographic factors that matter, and the developer doesn't want to do that, I think we could say, therefore, the method scores low on validity, just as we would if it were on clinical characteristics.

CHAIR NERENZ: Yes, Dave here. Let me just, I'll channel Larry here, but he can certainly jump in.

I think the connection to reliability was this idea that, if you're dealing with signal-to-noise statistics and then, because you didn't do acceptable risk adjustment, some of the apparent signal is based on things other than

quality, then the lack of adjustment means that the reliability statistics are inflated and, therefore, questionable.

MEMBER STOTO: Okay. Thanks. This is Mike again.

That clarifies it, and I agree with that, but I still think that what I said holds. If we think that it should be risk-adjusted for whatever characteristics, and if it's not, that's a reliability -- excuse me -- that's a validity issue.

CHAIR CELLA: I'm just going to try to see. Jack had his hand up, but, apparently, he's on mute. I don't know if he can speak. Jack, why don't you give it a try?

MEMBER NEEDLEMAN: Okay. Can you hear me?

CHAIR CELLA: Yes, I can hear you.

Let me add one thing just so people know. I believe that Larry is not on the call right now. He had to take an hour off for a COVID situation, but he'll be back later, if we

want to bring him into this.

But go ahead, Jack, and then, it will be Patrick and Paul.

MEMBER NEEDLEMAN: Yeah, just so we have consistency and we're scoring things, I think we want to decide whether the problem with risk adjustment is a reliability problem or a validity problem. I tend to think of it as a validity problem because the core of our measures for a lot of these are observed to expected. And the expected is driven by the risk adjustment. If the risk adjustment model is wrong, not doing what we think it's supposed to be doing of getting an accurate assessment, a valid assessment of expected, then the measure is not valid.

CHAIR CELLA: Yes, I suspect, based upon the setup of this question on the slide, that NQF agrees with that because they did say that this would not be a basis for voting low/insufficient on validity. And I think if they saw it as a reliability issue, they would

have probably said reliability. I'll let them speak for themselves, but my guess is that that's probably where NQF is on this question as well.

MEMBER NEEDLEMAN: Well, then, can we get some clarification? If we're supposed to be assessing validity, and we think poor risk adjustment is making the measure invalid -- we haven't got the right expected count -- are we supposed to nonetheless accept the measure on validity, waive that concern? What concerns do you want the next committee to do, put that in? How do they want us to handle this? I'm happy to handle it in any way I'm told to.

MS. WILBON: Right. This is Ashlie.

I think there's a couple of things that -- can you guys hear me okay?

CHAIR CELLA: Yep, yes.

MS. WILBON: Okay. I think there's a couple of things that would be potentially on the table for us. I certainly sense the frustration, and I do want to also reiterate that we do pass on that discussion. Each measure that occurs at

the Methods Panel level is summarized and provided to the standing committee in relatively great detail. We give them the preliminary analysis, the combined PAs. We give them a summary and their worksheets when they evaluate it. And so, those issues are passed on. I think, certainly, we could maybe find some ways internally, as staff, to make the issues I think a little more clear and stand out a little bit more.

The other thing is that we had this kind of moratorium -- I don't know if moratorium is the word -- but kind of prohibition in place in terms of voting down the measure before we had the policy in place where standing committees could pull measures for discussion, even if they did not pass. And so, something that could potentially be on the table is, you know, we could still allow the Methods Panel to vote as they have been voting, as they would like to vote, and in alignment with what they're seeing in the submission. And so, the signal about the

validity would be in the vote, but, then, standing committees would have the option to pull a measure for that reason for discussion. So, that's something we could, you know, it would potentially be in that bucket of measures that could be pulled for further discussion.

And at that point in time, we would certainly share the details about why the Methods Panel felt that that measure did not pass validity, if it was something specific to the risk adjustment or the inclusion or not inclusion of social factors. So, that's something we could consider.

Again, it's one of those process issues, but I think there may be some things we can do on our end. But, certainly, if there's other guidance that the Methods Panel has for developers, in particular, or for standing committees as they're evaluating it, I think that will be helpful as well.

CHAIR CELLA: This is Dave C. I'm just going to, I think, repeat what you said

there, Ashlie, but make sure that my repetition doesn't distort what you're saying.

I think you're indicating that NQF will consider a change in this position that not including social factors in a risk model should not be a basis for low/insufficient validity. And the reason you can consider that change now is that, between the time this policy was written and today, you've allowed for standing committees to pull measures back in for discussion. So, they would be able to pull in a measure application that has been deemed low/insufficient on the basis of social risk adjustment, or any risk adjustment, for that matter. The fact that they can pull that in would enable this committee to actually go ahead and make that vote, which I think gets at Dave N.'s concern about not knowing how to signal this concern other than putting a comment that might not be seen. Is that right?

MS. CLARK: This is Apryl. Yes. So, hi. This is Apryl. I just wanted to sort of jump in here.

I'm sort of new to the quality measurement side, but certainly I've been working on, I've been here since the social risk trial started. So, I appreciate the background.

I think I would want to make sure that we have the social risk trial actually being funded. We're actually looking at measures that come in. You have the conceptual basis and how we're sort of thinking about that. And before we make any changes, I'd want to make sure that we're kind of tying the work that you guys do to the work that that's doing and we can come out with sort of an organizational-wide kind of policy on social risk.

So, I'm not saying that we wouldn't necessarily consider having the Committee -- you know, as Ashlie mentioned, but I'd want to make sure that we're sort of thinking about it a little bit more broadly across the endorsement portfolio, tying it into the current work that we're doing, before we would make any policy changes. But I understand the frustration, as

Ashlie mentioned, about being able to do that. But I just want try to bring that other work in, make sure that we're thinking about it, you know, sort of holistically.

CHAIR CELLA: No, that's helpful. So, Apryl, I was careful to say that consider as opposed to making a policy change here on a phone call.

MS. CLARK: Yes.

CHAIR CELLA: So, would you say that that is accurate, that sort of a reconsideration of this could be done in the context of the larger picture, as you say?

MS. CLARK: Yeah. I mean, I think that there is lots of discussion going on all around the table --

CHAIR CELLA: Right.

MS. CLARK: -- including with our standing committee, about this issue.

CHAIR CELLA: Yes.

MS. CLARK: So, I'll make that part of the discussion about that.

CHAIR CELLA: Okay. So, thank you. That's very helpful.

So, with that in mind, are we now on Patrick? Or Paul?

MEMBER ROMANO: Yes.

CHAIR CELLA: Patrick. Go ahead, Patrick. And then Paul, and then Gene.

MEMBER ROMANO: Yes, I think that I'm putting this discussion in the context of the refresher that Ashlie gave us a few hours ago when I was just waking up about the validity and how we assess validity, and the fact that we can take into consideration the performance characteristics of a risk-adjustment model, including its discrimination, its calibration, et cetera, but we're not supposed to take into consideration presence or absence of any specific risk factor. And, of course, as we're now talking about social determinants, that fits into that framework clearly.

But, obviously, this is a bit of a funny distinction, right? Because we know that,

let's say that we see from the evidence that has been presented that the model is miscalibrated with respect to low-income people, right? I mean, the natural consequence of excluding social determinants from a model where those determinants belong is that the model will now be miscalibrated for those individuals. So, the absence of those factors from the risk model becomes a calibration issue, and then, it becomes fair game for us to discuss, as it were.

In other cases, a risk model may appear to perform well. I mean, the C-statistic may be, let's say, .75. But we know from previous published literature that the C-statistic ought to be .85. And if they had adjusted for other things that they didn't adjust for, then it probably would have been .85.

So, again, that leads us to say, well, even though .75 might be considered adequate, we know that it's not where it should be, and this means that there's residual confounding or omitted variable bias. And we have a pretty good

sense what those omitted variables are, even though we're not the clinical specialists who are on the standing committee, and therefore, empowered to comment on the specific things that should be in the model.

So, I'm just saying that this idea of evaluating the overall performance of the model, including particularly its calibration as well as its discrimination, is not completely separable from this issue of what's in the model.

CHAIR CELLA: Thank you, Patrick.

Paul?

MEMBER KURLANSKY: A few comments, though, some of which are going to echo things that have already been said.

But I don't think we need to say definitively that this is an issue of either validity or reliability. I think Larry's analysis is probably pretty sound that a very inadequate risk model can skew your assessment of reliability. I mean, I have to agree that it is fundamentally an issue of validity, but it can

certainly impact reliability. So, I don't think we need to say that it is purely an issue of one or the other. I think we can just acknowledge that it impacts both.

Having said that, I guess I would strongly hope for a potential change in policy because, as other people have said, if we assess a risk model and find that it's methodologically flawed and inadequate, so if we find that it's methodologically flawed and inadequate because of this particular issue, then we can't comment on it or we can't use that to vote that it is low validity. But if we find a methodologic issue, that it's inadequate based on some other parameter, we can. It puts us in a very compromised and untenable position, just as a methods group.

So, I think that there is a certain ambivalence right now in the policy to where it has to be included and assessed, but it can't be used as a criterion. I think that really puts us in a very odd position.

The final comment is, having said that, this gets back a little bit to the issue that we talked about before as to what the metrics are going to be used for. For example, if you elected not to put in these factors because you were going to use this metric in order to identify those sites which needed more assistance and more support and more finances, then you might specifically want to leave this metric out -- I mean, excuse me -- you may want to leave these parameters out for that particular reason, because you want to identify sort of underserved sites that need help. Then, it might be a very compelling reason to leave it out.

But if we can't know or determine exactly how the metric is going to be used, then I think we have to go back to the sort of maximal criteria. And based on maximal criteria, I think if the risk model is inadequate, it's inadequate.

CHAIR CELLA: Thank you, Paul.

Gene?

MEMBER NUCCIO: Yes. Hi. In the

interest of complete disclosure, I was a member of the NQF panel that Dave Nerenz chaired a few years ago on sociodemographic risk factors. And so, I come to this with a certain bias.

Kind of working from the bottom up in terms of the questions on the screen, not all risk factors have an equal impact on an outcome or a provider's ability to achieve an outcome with a patient. And as the number of risk factors increases -- and I had examples this time around where there were in excess of a hundred risk factors that were used in the equation for the prediction equation -- obviously, no single risk factor is going to make that much difference.

So, I think the rationale provided by some of the developers that, well, it only accounted for 1 percent of the difference, and so that's not much, so we're going to ignore it -- one of these sociodemographic risk factors -- is bogus. Because when you put a hundred risk factors into an equation, individually, there

isn't going to be a lot of change with any individual risk factor, adding it in. Stepwise, it might show something different.

The second question that's asked is the decision to include or exclude in the provider's control. I call this is the developer. So, in many cases the developers did a really excellent job of showing that, in fact, there is a relationship between a patient risk factor of which the patient -- or the provider has no control over, such as maybe the availability of health care institutions in this area or nutrition, or whatever, but then simply decides not to do it.

Well, if the developer can simply ignore the suggestion -- I think it's stronger than a suggestion -- the directive from the NQF that sociodemographic risk factors be included in a prediction model for risk adjustment, then, you know, what teeth is there in terms of having them acknowledge the fact that hospitals, for example, that deal with patients who are primarily

homeless are going to have poor risk factors if you're not adjusting for whether or not they have a permanent residence? Now we're not asking them to correct the problem of homelessness, but we are simply saying that the outcomes that the provider can achieve is directly affected by these kinds of risk factors.

And finally, I agree with the previous presenter or caller where they said it is not validity or reliability; it's both, primarily, in my mind, validity. And if we're asking if these measures are being used to make distinctions among providers, not just in terms of the quality of their work, but also in terms of whether or not they get rewarded monetarily -- and there's a lot of value-based purchasing measures out there -- then we need to enforce in some way that, when you demonstrate that there is a relationship between the sociodemographic risk factor and an outcome, that those need to be included just as the clinical risk factors would expect to be included.

I'm sorry I took so long.

CHAIR CELLA: No problem, Gene. Great points. Thank you.

Christie, you've got your hand up. Go ahead, Christie.

You may be on mute, Christie. Can you try to unmute yourself?

(No response.)

MS. WILBON: We'll work with Christie to make sure she's connected.

CHAIR CELLA: Yeah. I think we'll do something offline here, some on-the-side work to see if she's muted remotely, and we'll come back to you, Christie.

Jennifer has her hand up. Go ahead, Jennifer.

MEMBER PERLOFF: Okay. Great. I have multiple phones ringing here.

One concern I have with some of the sociodemographic risk factors is that they often require bringing in external or new data not necessarily related to the core data source, like

claims or surveys. And we rarely see what missing data gets introduced into the mix because of these factors being added. So, are there missing data elements? Does the assessment sample size go down when you try these measures? That's just something I feel like we never get good visibility on, and maybe it's a non-issue sort of in this space, but a little bit tangential, I just wanted to raise it.

CHAIR CELLA: Okay. Thank you.

I guess Christie's going to be trying to dial back in to a number that allows us to hear her.

Any other comments?

CHAIR NERENZ: Yes, Dave N. here. Just a quick response to Apryl.

At least in my framing of this issue and the questions I raised, I wasn't seeking in any way to change broad NQF policy or to get out of line with broader discussions of policy about social risk factors. It was really about how to implement the policy currently in place, where

developers are asked under certain criteria to include social risk factors if a set of criteria are met. And our question is, as the Methods Panel, what action can we take and what decisions can we make if the treatment isn't satisfactory according to the policies that currently exist? So, again, not pushing for any change that's out of line.

CHAIR CELLA: This is Dave C.

I don't know if Ashlie or Apryl want to say anything to that.

But, if I'm not mistaken, at least one of the seven submissions we'll talk about, this will come up in that context. So, even though it's good to know that NQF is going to take this up under kind of a more general set of issues that are under review and discussion, we will be hoping for some guidance during this meeting, specifically when these come up, when this comes up again in the specific measure context.

I still haven't seen Christie join in. So, we'll go to Eric and, then, Jack.

Are you there, Christie?

MEMBER TEIGLAND: Can you hear me now? Yes?

CHAIR CELLA: Yes, I can hear you now. So, go ahead, Christie. You're up next.

MEMBER TEIGLAND: All right. Yes, I had to call in. I could hear, but I couldn't speak, I guess.

Yes, I just wanted to reinforce everything that has been just said. This has been a thorn in my side since we started this process and it seems to be getting worse, not better.

And I just wanted to bring it home with a real example. Well, there are many in the measures I evaluated this round. But there were some cost measures, and, yes, they did a great job, as you said, laying out the conceptual basis. They found that the social risk factors were significant.

And when they presented the data, this time they actually showed what the difference

was. These are cost measures. So, the cost for a dual-eligible patient in a skilled nursing facility is 18 percent higher across. For a non-white in a skilled nursing facility it's 25 percent higher. But, then, they said, yes, we already have 120 variables in the model and this doesn't improve the model fit.

What does that say about those skilled nursing facilities operating on very tight margins who have 14 to 25 percent higher costs, who are treating patients with these social risk factors? Is it going to hurt the quality of care or is it really going to impact access to care for those people?

And the variables were significant even in the model. It wasn't even like, oh, we added them in at the end. And, of course, they didn't add any -- you know, they weren't significant. They were significant in the model.

So, there's something wrong with the model, right, when you have that big of a variance in cost for these groups, but you're not

finding that in the risk-adjustment model. So, there's something wrong here.

There was absolutely zero rationale, as far as I could tell, to not include them, but you're right, they just said, but we're not. We decided not to. So, I think we need to do something about those kinds of situations.

CHAIR CELLA: Well, that really brings it home. Is that one you're referring to one of the seven we'll be discussing?

MEMBER TEIGLAND: I don't think we're going to discuss this measure. It's the PAC Smith measure. I think they passed. Yeah, I think they passed. These measures passed because they had decent validity/reliability, which kills you to make those judgments when you know they are not being appropriately specified.

CHAIR CELLA: That really does bring it home, Christie. Thank you.

MEMBER TEIGLAND: Yes, yes.

CHAIR CELLA: Eric?

MEMBER WEINHANDL: Hi. This topic

absolutely tortures me. And part of it is because I've got training in epi and I've got training in biostats. And the two parts of my brain don't really totally agree on this one.

(Laughter.)

MEMBER WEINHANDL: And then, complicating the fact is, you know, that when I think about these social risk factors, definitely, my mind always goes toward the denominator of patients who are Medicare-eligible or Medicare-enrolled. The simple binary factor of just the previous example: are you also enrolled in Medicaid, yes or no? And so, that's the easiest one for me to think about. Obviously, there's lots of other social risk factors that could be used, and some of them are measured on a continuous gradient, and the methodological challenges become more severe.

I guess a few thoughts. One, I want not to forget is just reacting to the text on the screen. And that is that question of the word control. I guess it's the second-to-last bullet

point.

I find that the notion of control is pretty context-dependent. And what I mean is, imagine that you're concerned about something like food supply or about home security. Home security may be a poorer example than food supply. But you can imagine how there are, depending on which provider is being evaluated in this measure -- maybe it's a health care provider or maybe it's a Medicare Advantage plan sponsor -- there are genuinely different statutory restrictions about what can be controlled in the home for a patient or for a person, depending on who the actor is that is being evaluated.

And so, there's a part of me that says -- and I don't know if it's the word control or in in the provider's control. I don't recall. I have read the report from the NQF before, but I don't recall if that specific language is part of the policy or invoked in the report.

I think that it's tough to ask, say, the Scientific Methods Panel, to opine, or even a

measure developer maybe, to opine on whether something ought to be in control or not, unless you have a really detailed knowledge of what the legal restrictions are for what is in control. I mean, Medicare Advantage plan sponsors, by virtue of the 21st Century Cures Act, have the ability to do certain things now with respect to social risk factors and food security that they didn't have the province for before. So, you can imagine it then becomes complicated.

I'm definitely receptive -- so, that's kind of the policy part of me. The biostats part of me says, if you know that, among your list of risk factors that explain your outcome, dual-eligibility is a prognostic factor, then it's kind of a black-and-white decision for me. It should be in the model. And I suppose that my judgment there is partly because of this lack of clarity about what's in control versus what's not in control.

I mean, it's easy to talk about a social risk factor and pretend that we think that

this is something that's exogenous to the system and that, you know, you can make a principled argument that you should be treating people who are poor the same way that you treat people who are not poor. But couldn't you make the same argument rather easily about a lot of comorbid conditions? If you have a drug therapy, would you make a distinction between heart failure, yes or no; diabetes, yes or no?

I find that when I think very critically about this question of whether there's something truly distinctive about social risk factors, I often find myself eventually asking, well, is there actually anything different or is the same ethical decision in play for comorbid risk factor as for a social risk factor? So, then, the distinction begins to wash away.

And then, I guess that, on a related note to all of that, it is that, when it comes to the volume of risk factors in a risk-adjustment scheme, I mean, no doubt, with administrative databases, you're going to consistently find the

poverty of associated dichotomous, continuous gradients -- it's going to be associated with poorer outcomes.

But I often find that, to some degree, when you see that independent adjusted association of, say, dual-eligibility with a poor outcome, there's often a connection between the strength of that association and how many other medical factors are in the model. And sometimes I think that we concentrate on the social risk factor when, actually, it's a proxy for poor adjustment for a lot of other medical risk factors.

And that's why I bring it up at the exact same time as talking about the distinction between social and medical risk factors. Because, from a statistical perspective, I'm not sure that we're looking at anything other than a large collection of very interrelated correlated factors. So, that's the biostat part of me that's speaking.

Finally, the epi part, and then, I'll

wrap up. One of the things I'm struck by is that, if we were to accept that we should include social risk factors if we know that there are good data that support an association between them and outcomes, would it be reasonable or, is it overstepping, for this evaluation process of measures to encourage that, when measures are reported, they aren't simply reported en masse with an adjustment for social risk factors, but that the measure is also reported in those who are poor and not poor, those who are dual and not dual? So then you have an opportunity to see not only the risk-adjusted perspective.

You know, given the world as it is, how is the provider doing? But also understanding for that provider what is their difference in all terms between poor and not poor. So that you can begin to understand whether that delta is actually unique for that provider versus other providers.

CHAIR CELLA: It's kind of like reporting actual observed means and modeled means

in a patient.

MEMBER WEINHANDL: Yes, yes.

CHAIR CELLA: Yes.

CHAIR NERENZ: Dave N. here.

That actually was one of the specific recommendations of the 2014 panel report.

CHAIR CELLA: To report both?

CHAIR NERENZ: Yeah. If it makes sense on all the criteria to do adjustment and use adjusted measures, that's fine; that's good. And that was a change in NQF policy.

But also then, a parallel recommendation listed right in there is to encourage reporting of what's referred to as stratified rates. So, it's basically what was just said, that, depending on what the variable is and what the relevant categories are, report out referring to that stratified fashion. And I thought it was feasible because of some of these issues and a few other considerations. But, as a concept, yes, it's desirable.

Now that gets us into this territory

of use and reporting and who mandates the reporting, and that sort of thing, but we did agree clearly as a concept it's an important and good idea.

CHAIR CELLA: Eric, did the epidemiologist part of you finish?

MEMBER WEINHANDL: Oh yeah, that was my last part. Thank you.

CHAIR CELLA: Okay. We heard from three Erics. You said epi and biostats, but we also heard from the policy Eric. So, thank you. It was a great expose.

Now Jack had his hand up, but I don't see it up now, would be next, if you do want to speak. And then, Larry is on.

So, Larry, welcome back.

MEMBER NEEDLEMAN: Yes, my hand is up.

CHAIR CELLA: Okay.

MEMBER NEEDLEMAN: My hand is up and somehow the screen seems to keep lowering it.

CHAIR CELLA: Somebody doesn't want you to talk, Jack. It must be that.

(Laughter.)

MEMBER NEEDLEMAN: Right. Okay. Very rich discussion, and I think a discussion of risk adjustment in general dealing with issues of what's downstream, what's under the control of providers, is particularly to some of the cost measures which look downstream three months, six months, nine months, is valid. But I'm going to stick with social risk adjustment for the moment.

And I think there are a couple of things we have been seeing in these measures. Often, the reason for we're not going to add them is it doesn't add very much to the predictions. The r-squared doesn't change very much. The relative rankings of folks don't change very much. And by very much, what I really mean is what we're told is the r-squared changes virtually not at all and the rankings change virtually not at all. So, we're basically saying, if we throw these in, it's de minimis.

And I think it's useful to reflect a little bit on why that's the case and how it's

affecting the rankings it provides. One of the reasons is some of the things that the social risk factors do are downstream from and included in the models through other downstream variables. So, we've got mediated relationships. The marginal contribution of the risk factors to the outcomes are smaller than the total contribution of the risk factors to the total through all the different pathways. And that, I think, is one of the reasons why Christie commented about being higher, and so forth. That's, I think, one of the reasons why.

But the second reason is we are only looking at Medicare data. This issue keeps coming up in the CMS measures and we're only looking at Medicare data. And what we know about the facilities, particularly the facilities that treat high-volume either minority patients or low-SES patients is they tend to have smaller volumes of Medicare.

So, we're looking at a small contribution, a small volume in these providers,

adding to our uncertainty, but, also, you're going to see higher volumes of Medicare patients, including some in not particularly low-SES-serving providers, where some of the low-SES folks there, you're going to see the balance there.

The concern that drives a lot of the are we being unfair to low-SES-serving providers is about their capacity to provide the kinds of services and meet the kinds of needs that their patient loads have. But an awful lot of the patients that we're concerned about are not Medicare.

So, we've got this real imbalance between a small number of Medicare patients in those facilities, and they're not the only ones whose care we're worried about. And the Medicare quality is focused on just the Medicare patients. So, we've got this small contribution to r-squared, the small changes in ranking being driven by the fact that we also have smaller numbers of Medicare patients in those low-SES-

serving facilities. And I've been told this by a couple of folks who run these. Given the small number of Medicare, given that they're often paid by Medicaid lump sums or per diem, they don't do as a good of job coding, which means, at the same time you get some of the SES coding there, you may have lower risk adjustment scores on the other factors that influence the risk adjustment because all their comorbidities are not being coded on.

So, we've got a broader issue of, should we be just including this as a matter of principle? Should we be not including it because it doesn't change the r-squared very much? And that, I think, has been some of the context in which we've seen these measures discussed by the developers where we would have to figure out what we want to do about it.

CHAIR CELLA: Thanks, Jack.

Larry, you're up.

MEMBER GLANCE: Great. I'm sorry I missed a good 30 minutes of this discussion, but

I'll make a couple of comments, and I'll try to be succinct.

The first one is that, at some level whether or not to include SES is not so much a statistical question; it's a philosophical question. So, if you believe that socially disadvantaged patients are more likely to have worse outcomes and to incur higher costs, and you believe that institutions/hospitals that care for these people are being potentially financially disadvantaged with these performance measures that do not account for SES, then you end up hurting those more vulnerable institutions that are taking care of the most vulnerable populations. Then philosophically, I think it makes sense that, when that sociodemographic information is available, that it should be included in the risk adjustment, period. So, that's the first point.

The second point is that I agree that the standard approach that CMS and some of the other measure developers have used is they'll put

SES in the model and they'll look at overall model performance, and they say, it doesn't change. And then, what they'll do is, again, they'll put SES in the model, if they choose, SES will move out, and then, they'll compare hospital rankings overall and they'll say, it doesn't change. It doesn't matter. We're not doing it.

I think maybe a more pragmatic approach, and one that we could maybe discuss, would be to say, okay, you know, if you have 2500 hospitals, it probably isn't going to make much of a difference overall for that population of 2500 different hospitals. But if you were to limit your comparison to, say, safety-net hospitals, those hospitals that, say, have more than the top 10 percentile or quartile in terms of their proportion of socially disadvantaged patients, and then, see how much of an effect including SES has on the rankings for that particular group of hospitals. And then, if it has a significant impact, then you ought to include it.

The problem is I think it would be hard to come up with a cutoff for how much the rankings have to change. But I think if we could sort of maybe talk about that a little bit, that might be one way to get around that.

But, again, philosophically, there really is no cost to the measure developer to including sociodemographic risk factors if they're already in the data. And I think, although people have debated back and forth whether you should include them or not include them -- and I'm not going to get into that whole discussion -- I think most people, at least what I'm hearing, are falling into the camp that, yeah, you ought to include them.

So, why not, as a group, why shouldn't the NQF say, look, if you have that data element -- and certainly CMS does -- you ought to put it in the model, period?

Thank you.

CHAIR CELLA: What a rich discussion that's been. I don't see any -- oh, now I see

another -- I see Paul's up. So go ahead, Paul.

CHAIR NERENZ: And then Sherrie.

MEMBER KURLANSKY: There is a quantitative metric that you could use called the Net Reclassification Index, which does have certain parameters attached to it that you could actually see what would be the impact in terms of reclassification by including it in the model, which may be more meaningful than how it changes the C-statistic, which frequently is minimal, because the C-statistic deals with the overall population, but not necessarily with the impact on individual sites.

CHAIR CELLA: Elaborating on Larry's point. Thank you, Paul.

Sherrie? We can't hear you, Sherrie. Did you call into the 833 number? It seems to me you need to call the 833 number to be able to be heard.

Why don't we wait and see if Sherrie's going to maybe redial? She sent a note saying she's back on the call, but maybe came in through

the other number.

Quite a rich discussion. I see Christie and Paul have hands raised. I think those are recycled raised hands. So let's check.

Christie, do you have a comment?

MEMBER TEIGLAND: I just had one quick other comment. And that was that I also saw a couple of measures this cycle where they made just the opposite argument for including some clinical factors that didn't show statistical significance in the model, but they let them in purely for face validity, which we know has been done for a long time.

CHAIR CELLA: Yeah

MEMBER TEIGLAND: So I like to -- you know, when you talk out of both sides of your mouth kind of thing.

CHAIR CELLA: Well, I think it goes back to, I think it was Eric's point that, you know, at the end of the day, there's not much difference between social risk adjustment and clinical risk adjustment, and they are correlated

risk factors, at least from a statistical perspective. So it's hard to make the case that one's more important than the other.

Okay. I still don't see Sherrie back. She says she called the 833 number and "unmuted my line, but still can't be heard."

MS. WILBON: Sherrie, try hitting star-1. See if that gets you on.

CHAIR CELLA: I still don't hear her. I assume she heard you saying hit star-1.

Let's go to Daniel, and then we'll see if Sherrie has found her way in.

Go ahead, Dan.

MEMBER DEUTSCHER: Just a couple of notes. First, as a general perspective, I think at least from my perspective, I'd like to see developers first test for SES factors, and is actually currently expected. Now we do have sometimes -- I've seen, you know, with my real experience on this panel, some cases where they just say, "We don't have any SES data."

So I think we also need to address

this situation. So what do we do about that? Is that okay? Because that's a really easy way to get around that, just say, "We don't have the data." So are we expecting or will NQF be expecting in a certain amount of time for every measure developer to get SES data and test for it? That's one question.

Another thing I think is that, once an SES factor has been identified as significant, whether it contributes more or less to the predictive power of the model, I can't think of very good reasons not to include it in the model. And I just want to give an example.

There are studies that show that an SES factor or grade is significant, that it's no longer significant once treatments are adjusted for. So I think that's a case where developers should really be careful about including an SES factor that goes away once treatments are adjusted for.

And there could be multiple reasons for that. Maybe low-SES patients get worse

treatment for some reason that's not really dependent on the overall resources. Maybe it's just the selection of treatments that's different. And there are studies that show that.

But, in that case, I would expect the developers to provide evidence to support that, not just throw out a philosophical argument that's not supported by evidence.

Those were my two comments.

CHAIR CELLA: Thank you. Thank you, Daniel. We're going to check back with Sherrie again, who's going to try to speak. Go ahead, Sherrie.

MEMBER KAPLAN: Can you hear me now?

CHAIR CELLA: Yes, there you are.

MEMBER KAPLAN: Sorry about that, Dave. I think they entered me as a non-presenter.

I just wanted to back up and agree completely with Larry about the need to sort of -- the face validity question, if you will, of including socioeconomic status data where it's

available in the data source. And that is, there's evidence from a number of studies, some of them are older, that academic medical centers and public hospitals are disproportionately represented in the penalty arena for adjusted compensation by CMS.

So I think -- you know, and some of the argument is well, we don't have -- some of the case mix of variables should adjust for all the arguments that poor people are sicker, but, in fact, there's a lot of evidence that -- for example, the Area Deprivation Index in our local situation moves hospitals around in terms of the distribution. So I think that, with just a blanket you need empirical evidence collection, you know, does the empirical evidence support it, in this day and age probably isn't adequate.

I mean, there's a fair amount of opinion out there that this should and does matter just from a philosophical standpoint. And I think that warrants at least watch, you know,

look at the adjustment, see if it is costly, and there's probably no reason not to, if possible, do it.

CHAIR CELLA: This is Dave C. again. A quick note that we're being reminded, and I don't think this applies to Sherrie, in particular, but the use of a speakerphone is not encouraged because it's harder to get a good quality tape recording. So if you can avoid using a speakerphone when you're speaking, please do.

The other thing I wanted to say is that I think this has been a fascinating and very rich discussion from several people who have thought deeply about this topic.

And the note to NQF I think is that I think it's a fair summary to say that there is a lot more support for including social risk adjustment than for not including it, even if the delta in the r-squared is not substantially changed, although I don't think that's everyone's opinion, but I think I heard more in favor of

including it, at least on philosophical grounds and also even on statistical grounds in terms of having a common, you know, similar standing with clinical risk factor adjustment.

So I think it's a fair summary to say that the majority of the Committee does endorse NQF going back to look at this, as Apryl said, in a larger context and maybe letting us know at our next meeting where you are with that.

Understanding that Dave N. raised this not to start a revolution, but it did stimulate quite a discussion. And so with knowledge that Dave's motive was really to get guidance on how to respond now to these submissions, it does raise the larger policy issue. And it's good to know NQF is paying attention to it and will be revisiting this with the standing committees and internally.

Is that a fair summary?

MS. WILBON: Sure. Hi, Dave. It's Ashlie. I think that was a great summary.

And I did also just want to note that,

you know, I think there are things that we can be doing at NQF, to Dave's point about not starting a revolution, to make sure that both developers and committee members, and NQF staff, really understand what the current policy is under the trial.

And so I think we can do more on our end to make sure that that's communicated and to make sure that the policy is accurately kind of carried out and applied in the measure evaluations at the standing committee discussion as well. So thank you for that summary and agree -- we agree.

CHAIR CELLA: And I think -- you know, it's probably not appropriate for me to volunteer people, but, you know, you heard from at least a dozen people who have obviously thought a lot about this and can be helpful, I think, to you in going forward with your thinking about it.

In the meantime, I would ask everyone to keep this discussion and the earlier discussions from the morning in mind as we do go

through the submissions that we have to vote on as the meeting progresses.

So we're still -- sorry, go ahead.

MS. WILBON: I was going to say, Dave N., really quickly, did you already start an outline for this topic? I know you had mentioned that you had started taking some notes and had some ideas about a paper. Is there something that you already have in mind? And should folks reach out if they are interested in helping you with that?

CHAIR NERENZ: Yes, that's right. I was motivated just to put some things on paper as I was doing some of the reviews of measures that I thought needed some kind of discussion made about inadequate risk adjustment. And so maybe if it's just a process issue, if people want to let me know if they're interested in being part of the writing group, I can send around at least what I've got, getting started.

I wanted to hold it pending this discussion this morning to just see where people

are coming from and who seemed engaged in the issue, and if there was a general direction where we were going. So we can pick it up from there, if there are folks who are interested.

MS. WILBON: Thanks. That's helpful. Thank you.

CHAIR CELLA: Okay. Well, let's proceed on to the next topic on the agenda, if you can advance to the next slide. It's on cost, cost measure evaluation challenges.

And I think Jack and Jennifer were tagged for teeing this up and helping to set up a discussion. Jack and Jennifer, do you want to walk us through this?

MS. WILBON: Dave, I'm going to give just a quick tee-up, and then I'll have Jack and Jen, and actually Bijan, share some thoughts.

CHAIR CELLA: Sure.

MS. WILBON: They were part of a recent discussion with the Cost and Efficiency Standing Committee.

CHAIR CELLA: Perfect.

MS. WILBON: But I tagged Jack in particular because he emailed these issues. So definitely would appreciate his input on expanding on some of these questions that are here on the slide.

But I did just want to give a little bit of context. Obviously, cost measures come to the Scientific Methods Panel for evaluation. We got six this cycle. We are anticipating seeing at least a handful of measures every cycle going forward. I think CMS is working on developing several measures that will be submitted to NQF over the next several cycles. So this will be kind of a mainstay in our portfolio going forward.

And so I think it is certainly important to have some principles that underpin our evaluations for these to make sure there's a common understanding of how we're approaching the evaluation for these. And I think that the Cost and Efficiency Standing Committee is grappling with some of the same issues. Many of them are

economists and methodologists as well as clinicians.

And I think we, understanding that, we actually put together a call just a couple of weeks ago with them to discuss some of the challenges, particularly in evaluating validity, based on some of the concerns that were raised in the spring around the validity testing for some of those measures.

So we had, I think, a really great discussion with the Cost and Efficiency Committee. And in addition to the issues listed here, I just want to kind of offer an opportunity for Bijan and Jen and Jack to share some of their takeaways from that discussion and maybe summarize some of the issues here on the slide for us.

MEMBER PERLOFF: Great. Jack?

MEMBER NEEDLEMAN: Okay. So I'm just talking about the evaluating exclusions issue. We've got two bullet points on this page, just the evaluating exclusions issue.

We've seen something in the submissions that I do not recall seeing before. You know, we've got a number of these submissions where there's an index event, and then we're going to look at all the things that happened three months, six months, I think we've got one that's nine months, out. And sort of attribute that to the provider at the incident event. And I'm not arguing about that logic. That's a whole other subject.

But what they said in the specifications of a number of these is we have excluded certain downstream events that the committee, our Technical Expert Panel, and the CMS docs judge outside of the control of the incident provider. It might be -- one of them was planned readmission, which makes a lot of sense.

But the issue that I had, as a reviewer of these things, is I had no idea how the process of identifying the costs for exclusion as outside of control and/or otherwise

predictable were being done. Was there a systematic review of possible things? Did they say yes to this one and no to that one, and we did not see the noes? Did they -- is it an ad hoc process where somebody says well, I think this could happen and we ought to exclude it? And only the ones where somebody said this could happen, we ought to exclude it, sort of got on the agenda for discussion.

And none of that is clear in the documentation we've gotten. So it's hard to evaluate whether the exclusions look reasonable, but I don't know what they decided has not been excluded or how systematic that review is. So I don't know how good the exclusion process is, and that was the concern I was raising.

MEMBER PERLOFF: I would raise a different exclusion issue, sticking with this concept of the episode or bundle of care. Jack is mentioning sort of the exclusions within the construct, things that are not counted when you're thinking about the cost or the quality of

care.

But another type of exclusion is episodes themselves that get dropped from analysis. And we have seen measures where something like less than 10 percent of the originally-triggered cases remain in the final measure. And we see some where it's small, 5 or 10 percent dropout, and some where it's tremendous. And again, not much understanding of what is left on the cutting room floor when you have that degree of dramatic exclusions. So just a different take, the same theme.

CHAIR CELLA: This is Dave C. I mean, I confess that I was, you know, when I looked at the exclusions, I kind of had my clinical trialist hat on and thought in terms of, you know, did this exclusion seem reasonable based upon the kind of underlying rules that I might use for deciding whether or not an exclusion from eligibility in a clinical trial or excluding from an outcome determination in the trial made sense.

And so, you know, it kind of brought

me to that provider control issue, if it seemed like a cost that was legitimately -- either it wasn't a major cost or was something that was outside of the provider control. It seemed reasonable.

But I agree, I never saw it spelled out. I was using other sort of internal models.

MEMBER NEEDLEMAN: This is Jack again. If I can just -- I totally agree with you that the exclusions we've seen, to me as a non-clinician, have all made reasonable sense. It's more of how systematic is this review -- and how, you know, what didn't get excluded. Again, it's a matter of how the process is done, which -- already mentioned. The trials that we've seen all look reasonable.

CHAIR CELLA: Right. Yes, yes. So you're not actually concerned about a specific case per se, but the process.

And Paul has got his hand up.

MEMBER KURLANSKY: So echoing Dave's sort of clinical trials hat, I am just wondering

if thinking in terms like a CONSORT diagram, if maybe -- I didn't actually have to review any one of these. But it seems to me that if you had a concept of what was the total population that was potentially theoretically at risk, and then having them demonstrate each exclusion, what percentage or what number were there -- excluded with a reason, and then what was the final study, you know, in sort of the format of a CONSORT diagram which you would do for a clinical trial, it might, to Jack's point, it might make things a lot clearer as to what you're doing.

Then you could make a better judgment as to whether or not the exclusion seemed reasonable or not or even if the extent of the exclusion seemed reasonable. But the process would be a lot clearer.

CHAIR CELLA: Good suggestion. Any other comments or questions on the evaluating exclusions component of this?

Larry?

MEMBER GLANCE: So in just listening

to this discussion, and I did not evaluate any cost measures, when I saw on the agenda that we were going to be looking at exclusions, I assumed that what we were going to be looking at is if you have outliers in terms of cost and how to deal with those. Are we saying that extreme outliers of costs are being excluded from the risk adjustment model, or are they being excluded from the population -- the patient sample that's -- whose cost performance is being evaluated? One or both? Or none?

What are we exactly talking about with respect to the exclusions? Could you make it a little bit more concrete for me?

MEMBER PERLOFF: On the high-cost cases, what we tend to see is winsorization. The high-cost outliers aren't excluded. They're truncated and sort of capped at the 99th percentile or the 95th percentile. So that would tend to be the issue there. It doesn't drive the case to be dropped.

We're talking about, I think, two

kinds of exclusions. One, in counting up the costs of producing a certain kind of care, or measuring the kind of utilization, some kinds of utilization costs are just excluded on theoretical grounds.

We couldn't control the costs. Physicians should not be held responsible for a patient who has a car accident on day 60 after knee surgery. That really is beyond anyone's control and should be excluded.

CHAIR CELLA: Good example, yeah.

MEMBER PERLOFF: And so, yes, anyway, I think that's, Jack, if I've got it right, those are the internal exclusions.

MEMBER NEEDLEMAN: Yes, that's right. The measure -- the definition is exclusion of unrelated expenses. And in none of the cases was the car accident three months after the surgery included as one of the excluded cases, the excluded expenses, nor did it exclude the case.

So I'm looking. I've got to go back to the -- in my notes, I did not document what

the exclusions, examples of the exclusions, and I'd have to go back to the original submission to go see what they did.

One of the -- and the appendices that they reference were actually not in our documentation. So there we are. The language showed up in, among other measures, 3564, Medicare Spending per Beneficiary, post-acute measure for home health, as an example.

MEMBER PERLOFF: Another type of --

(Simultaneous speaking.)

MEMBER PERLOFF: -- go ahead, yeah.

CHAIR CELLA: And then Larry after that.

MEMBER PERLOFF: I just was going to say the other type of exclusion in the skilled nursing facility stage, for example, that concerns me is cases that get dropped. So let's say you have a skilled nursing facility stay, and there is a transfer to a new facility or some other kind of similar event. Those cases as a whole will be dropped.

And to me, a concern is whether that's introducing bias into the sample of cases that are left. Is it really a valid measure of skilled nursing stay costs, for example, if I've excluded half of all cases constructed using the logic?

So back to -- I think, Larry, you were asking was it high cost. There are other things besides cost that lead to cases being dropped, characteristics of the episode itself besides being an outlier.

CHAIR CELLA: Larry?

MEMBER PERLOFF: Sorry. I just wanted to --

CHAIR CELLA: Sure. Thanks. That was Jen, right?

MEMBER PERLOFF: Yeah, that was Jen.

CHAIR CELLA: Yeah. Okay. Thanks. Larry?

MEMBER GLANCE: And that was exactly the question that I had in terms of maybe some examples of where there was concern about the

cases being inappropriately excluded. You can deal with exclusions by just including them as risk factors. So, you know, transferring -- a hospital transfer in, instead of excluding those, you may just want to put it in as a risk factor in your risk adjustment model.

And what I'm hearing, it doesn't sound like what you're dealing with is all that different from exclusion in other types of models. Am I missing something? Other types of measures, rather?

MEMBER NEEDLEMAN: Yes. What you're missing from me is a concern that I have no idea how things get on the agenda for considering being excluded as an unrelated cost, and therefore, how systematic the process is or how random this is in terms of the validity of the measure.

If you're going to start saying we're going to exclude unrelated costs, then you need to tell me how you're making decisions to consider whether costs are related or unrelated.

And that's not in the documentation at all, and it's not there how systematic it is.

MEMBER GLANCE: Got it. Thank you.

MEMBER PERLOFF: I would raise a separate issue, which in cost measures there is tremendous variance, right? The outliers are extreme. And dropping cases really helps a measure with reliability and validity. If you narrow the set of cases to be more homogeneous and similar, you're going to have a better-performing cost measure in terms of the things we've been talking about throughout this day.

But, in the process of narrowing that scope, you are leaving a lot of cases behind. And there's -- my concern is little attention given to what's left out and what that remaining set truly represents.

If we have 17 percent of cases as SNF care in the country used to assess resource use for SNF care, are we really -- is that really a valid measure? This is a validity question for me in a lot of ways. So I just wanted to raise a

different point than the one that Jack's making, which is, also, obviously, important.

And I think other measures, obviously, have these problems as well, but, for me, cost measures are very challenging as freestanding, individual episodes of care. They don't lend themselves well to being standalone, highly reliable, highly valid measures.

CHAIR CELLA: This is Dave C. I'm sort of torn between whether we're requesting guidance or whether we're nominating ourselves to help with this algorithm, whether a CONSORT diagram or some sort of explicit request for an explicit articulation of the basis for exclusion.

MEMBER PERLOFF: One of my thoughts when we met with the standing committee was that it would be helpful to do a white paper on some of the unique features and challenges of reliability and validity of cost measures. So I was hoping to nominate just as sort of a white paper topic which could include diagrammatic examples like we were talking about as one

dimension.

Because I think we're raising issues that we would like developers to be aware of. I'm not sure there's a short-term action.

But, Jack? Bijan? Others?

MEMBER NEEDLEMAN: I would just like more documentation of the process by which costs or, you know, services are nominated as unrelated to care and how they're evaluated, and a little bit of whether a decision has been made to not exclude some costs, just to get a better feel for the process here. It's part of my assessment of whether the process for putting the measure together has been valid.

CHAIR CELLA: That is Dave C. again. I'm sorry if I missed this. Is there a standing committee for cost measures?

MS. WILBON: Yes, there is. This is Ashlie. Yes.

CHAIR CELLA: And Bijan and Jen and Jack are on it?

MS. WILBON: Bijan and Jack are. Jen

we invited as a guest of honor to our webinar a couple of weeks ago, yeah.

CHAIR CELLA: So maybe this Committee could weigh in on endorsing the proposal to that standing committee to please do that white paper. It sounds like it's more their purview than ours. But we would appreciate the guidance. Is that fair?

MS. WILBON: Hi. This is Ashlie. I actually think it's probably a crossover. I think there are some issues that they are grappling with methodologically that they actually would like guidance from the Methods Panel on.

Some were around, I think the main issue that we discussed on the webinar a couple of weeks ago was around validity testing, and particularly for claims-based cost measures and how, you know, developers should be thinking about picking comparators and whether or not there is any kind of systematic bias or error created by using another measure of comparing an

administrative claims-based cost measure to another claims-based cost measure to demonstrate validity, and kind of what types of validity they're actually demonstrating by using that approach.

So I think there's definitely some methodological questions that have come up at the committee level that I think they are seeking some broader guidance from a group like the Methods Panel.

So, Jen, please feel free to add or clarify my explanation if there's -- I'm sure you can probably explain it much more --

MEMBER PERLOFF: No, no, I think that's perfect.

CHAIR NERENZ: What do people think about maybe some kind of joint committee or subcommittee that would have representation from the Cost Standing Committee and the SMP? And identify what are the four or five issues and, if not a white paper, at least write down some of your recommendations that we can then discuss.

MS. WILBON: This is Ashlie. I think that would be great. And I certainly would be happy to help gather folks across those committees and bring you guys together, so that there's a group that's focused on that.

CHAIR NERENZ: Sorry about this. Jack and Jen and Bijan -- well, I guess you said Jack and Bijan are members of that standing committee?

MS. WILBON: Yes, that's correct.

MEMBER BORAH: That's right.

CHAIR NERENZ: I think one of you or both of you would be logical leaders of this because you have a foot in both committees, if you would be willing to, I don't know, chair such a subcommittee to get the issues down on paper and convene people that are willing to weigh in on the requests that they have of us and that we have of them. Because it sounds like the requests are going both ways. Is that reasonable?

MEMBER PERLOFF: Hear, hear.

MEMBER BORAH: Yeah, this is Bijan.

I'd be happy to.

CHAIR NERENZ: Well, that would be great.

MEMBER NEEDLEMAN: Sure.

CHAIR NERENZ: So, Ashlie, should we leave this with NQF staff to coordinate a gathering of, you know, I don't know what the number is, but six to eight people or so to come up with identifying the issues and then drafting some position?

MS. WILBON: Sure, sure. We can certainly work on getting folks together for that. No problem.

CHAIR NERENZ: And then we can decide later whether this is something that we can circulate through email discussion or for the next meeting, depending on its urgency.

MS. WILBON: Yes. That sounds good.

CHAIR NERENZ: Okay. I don't see any other hands up. Are there any other thoughts on the topic? Or were --

MS. WILBON: Oh, Dave, sorry, just

real quick, I just wanted to know if -- Jack, did you want to just provide, before we move on, because we're ahead of schedule and we can break early and give folks a longer break. I am sensitive to the start time of Measure Evaluation since we've already coordinated that with the developer. So I'd like to stick to our schedule for Measure Evaluation.

But, Jack, if you want to just quickly, since we have a few extra minutes, just explain the second bullet around the risk adjustment? And then if folks have questions, we can, but just understanding that we are -- that can be a subtopic of the paper, if folks were interested in having more in-depth discussion.

MEMBER NEEDLEMAN: Sure. The second bullet, oh, this is the tailoring issue. So what we've seen in particularly the CMS cost measures is -- and to some extent in some of the other measures as well -- is they have migrated to a highly stylized, fairly standard way of approaching the risk adjuster. You know,

hierarchical HCCs, some specific clinical services, some interactions there. But, then, they're applying -- and it's always a lookback, right?

Sometimes they include the diagnosis or the condition that creates the incident event, sometimes not. But the other elements of the risk adjuster are always a lookback based upon review of the full records from the patients. So we wind up with one of the sources of exclusion is we don't have a full year of data before the incident event.

But there's not a lot of tailoring. And one of the -- and you keep seeing this language when they're doing the risk adjustment on a million patients, or a hundred thousand patients or several hundred thousand patients, that we don't want to add interactions, and we don't want to do this or that, because we're concerned about overfitting.

And even their model with 200 variables is not overfit on 100,000 or 200,000

patients. There may be specific issues of low cell counts, but it doesn't look like that. And so there's a tendency to not want to do a lot of tailoring to the specific service that's being examined or the specific providers that are being examined. And the question is whether we should expect more tailoring of the risk adjuster to the patients and care circumstances.

And in the case of the Medicare spending per beneficiary, in on example, they have month-to-month estimates which they average, but they don't take into account the hit by the bus or the new cancer diagnosis during the period in which things are being located. They don't think about why did patients get into rehab and do we need to make sure we've got specific measures. So it's kind of the variance in rehab, how people get into rehab, that leads to differences in cost downstream from the initial rehab circumstances.

And that's the question of how much tailoring we should be expecting versus just

saying they're using the standard risk adjuster. We know this. Sometimes it explains 30 percent of the variance, and sometimes it explains 11, and we shouldn't be worrying about what they could do to push the 11 percent of variance up to a higher number.

Those were the issues that sort of occurred to me as I was looking across the measures we've been getting as opposed to individual specific measures.

CHAIR CELLA: Thanks, Jack. I see that Gene raised his hand, maybe even before you started speaking. And then Eric.

MEMBER NUCCIO: Yes, I was just concerned that we're not going to get a chance to talk about risk adjustment. The problem that I saw, and I was part of the cost measure review team, was that there appeared to be a fair amount of what I would call cut and paste that went on across the four measures, the mean spending per beneficiary, that were presented. And perhaps the problem is simply that the developer was not

very careful in terms of this cut-and-paste methodology.

The one that really upset me had to do with home health, the mean spending per beneficiary for home health care. In that description, they say that the data for the patient is being taken from the MDS instrument. I can assure you, having spent 20 years doing home health analysis, that no patient in home health has been evaluated using the MDS. They use the OASIS instrument. And the items are not the same on those two instruments.

And so to claim that the model that they're using for the home health mean spending per beneficiary that takes into consideration the clinical condition of the home health patient using MDS is completely bogus. It just can't happen because those data don't exist. So that obviously upset me.

The other piece that I was concerned about, and it didn't have to do specifically with the risk adjustment, was it was not clear what

the meaning of the "mean spending per beneficiary value" was supposed to be. They, at the end of their presentation, showed that the data that they generated using their little models were normally distributed. And it looked, you know, like a typical normal distribution.

Yet, when they were going to be reporting the results, the results were U-shaped. That is, the lowest percentage of patients or units of performance, whether it's a SNF or a home health or an IRF or a long-term care facility, were U-shaped. That is, there were fewest people, percentage of agencies or facilities that were in the middle, and the largest percentages were either below the normal, below expectations or above expectations.

And I found it very confusing because how this setup -- what the criteria that were used to transform the normal distribution into a U-shaped distribution were not quite clear. And I'll end it there.

CHAIR CELLA: Thanks, Gene. Eric?

MEMBER WEINHANDL: Hi. This is an interesting topic. So I was on the subgroup that did not, to my knowledge or memory, did not evaluate any cost measures. So I'm definitely not speaking to anything that's in front of the panel this cycle, at least in front of my subgroup.

I will say that this has been an interesting thing I've observed in watching some of these Medicare cost measures being developed, and not so much through the NQF forum, but, you know, particularly through a variety of the payment mechanisms and models that have been proposed, especially from the Medicare and Medicaid Innovation Center, CMMI.

Historically, I've been accustomed to seeing a lot of cost measures being developed by academic groups that are working in contract with the government, with CMS, and so they might be Yale, might be University of Michigan. Some of these big contractors do a lot of measure development.

Typically speaking, my impression has always been that they are tailoring to risk adjustment schemes, to the disease stage, or to the application that they are evaluating. What I have noticed with the Center for Innovation is that they are, generally speaking, and I'm making a lot of generalizations, they're gravitating towards this blanket use of the HCC risk score.

And that -- the HCC risk score has got different components, right? There's a component for a general Medicare beneficiary. There's one for ESRD. There are ones for people who are a full year prior to eligibility, less than a full year. So there's some level of tailoring to different subsets or phenotypes.

But the bottom line is that they're taking this generalized risk score, which was developed for evaluating cost expectations, and it was developed on the fee-for-service data and is typically used for risk-adjusted Medicare Advantage data. But then they're using it in all sorts of applications from general to specific.

And interestingly enough, I even encountered an example specific to my disease -- dialysis, where the Innovation Center had actually proposed a payment model last summer where the outcome was home dialysis utilization. So we're actually talking about a medical outcome being risk adjusted with the HCC risk score. And I think to myself, well, now this is really daring because this isn't even economic outcome anymore.

It seems to me that the question, from my perspective, comes down to a validity question. And so I don't mean to become overly polemical about this. You've got a risk score that has been developed for a specific application, say the HCC risk score. It can be used, as any covariate can be used, in a variety of analyses and risk adjustment models.

You know, the child risk score, the Elixhauser scores, or adaptations of Elixhauser, are used in widespread scenarios. We know that those scores weren't necessarily perfectly

intended for each specific application, but they're used widely.

So I don't mean to say that this should be treated any differently, but I do think that it comes down to a question of validity. Does this risk score that was not tailored for your application provide appropriate resolution? And to my mind, the answer is yes or no, depending on what the nature of the economic outcome is.

If it's a general outcome, like total all cost spending across Parts A and B, I'm okay with it in general. You know, it would be good to see evidence before I say I'm okay specifically, but I'm okay in general.

When it comes to very specific instances, like episode payments of care, now I think the stakes are higher, and I think that you do want to have some awareness of what was happening before that episode began, what triggered that episode.

And my concern then, because of these

generalized risk scores not being tailored to the application, are going to explain a lot less of the variation. And as a result, the measure is going to be a lot less valid.

So I think it comes down to, if I were to really boil it down, for me, it would be, yes, we should expect tailoring if we're evaluating a very sort of short-term, highly-specific kind of cost endpoint. If it's all costs over a one-year time period, then I'm expecting less tailoring.

CHAIR CELLA: Thank you, Eric. Christie?

MEMBER TEIGLAND: Yeah, this is a measure that we are going to discuss tomorrow that didn't pass. But I just want to point out inconsistency in using HCC risk models and all of the ICD codes in those models to risk adjust this measure, but then because they wanted to apply it to both Medicare fee-for-service and Medicare Advantage, they elected to only use diagnoses from the discharge claim for a patient, which we all know is in no way going to document every

chronic condition that patient has.

And then it wasn't even clear if they're actually using history for Medicare fee-for-service patients because they have it, but they're not using it for Medicare Advantage because they don't have it, because they included some adjustment factors for Medicare Advantage.

But just the whole point of you developed the model using this huge set of data that assumes that you're going to be accounting for chronic conditions that patients have when you risk adjust, but then when you limit that to whatever diagnoses that were deemed important on that discharge claim, that patient could very well have multiple other chronic conditions that aren't captured. And so then, you know, the results are going to be highly skewed.

So, yeah, the measure didn't pass, but just this whole approach of using different data to develop the model, the risk-adjusted model, than you're using to actually calculate the measure seems completely inappropriate. So there

should be some kind of, yeah, rule against that. Maybe there is.

CHAIR CELLA: Well, thank you, Christie. I don't see other hands up. Wait. Larry just raised his hand. Go ahead, Larry.

MEMBER GLANCE: Just a really quick follow-up question to Jack and Christie. When they're using a CMS HCC model in their cost models, and I understand that it was developed in a different dataset, are they estimating new coefficients for the CMS HCC models for the individual categories or are they re-estimating a single coefficient for the entire score based on the old coefficients? I mean, they must be doing some kind of customization. Can you --

MEMBER NEEDLEMAN: Yes. The customization is in terms of whether they're including or excluding certain covariates. They're then taking the data, and they're running the regressions with all the covariates that they're including. And you get covariates, you know, you get measure-specific coefficients for

each measure. So the risk adjustment is being done there.

MEMBER GLANCE: Yeah, so they are customizing the model. And that's a fairly standard approach. I mean, certainly, the CMS HCC model is much, much richer than more traditional models like the Elixhauser comorbidity algorithm that's used repeatedly in the outcomes literature.

So can you explain to me why you feel that their approach is not sufficient? I mean, it's a very, very rich model with lots of different --

MEMBER PERLOFF: Yes. It's because the model was developed, I think appropriately, as you said, using all of the ICD-10s for the past year of history for that patient. But then when they calculate the measure, in this case they were only using ICDs from a discharge, from an inpatient claim, from an inpatient discharge claim. That in no way is going to capture all the ICD diagnoses for that patient for the past

12 months, right? That was the biggest problem --

MEMBER GLANCE: The interesting thing is that the CMS HCC model goes back and uses the carrier data as well. So it uses not just inpatient, but also outpatient --

MEMBER PERLOFF: That's right. That's right. That's right. It uses a history, a one-year history, of claims to calculate the HCC risk score, yes.

MEMBER GLANCE: So what is the major limitation of using the CMS HCC model for cost --

MEMBER NEEDLEMAN: Okay. So we looked at measures in this last cycle, and I haven't got the full list in front of me, and I apologize, by name, but we looked at measures for home health, patients who had been admitted to home health, patients who had been admitted to SNF, patients who had been admitted to rehab. And there were a couple of others.

They don't always include the incident diagnosis that brought the patient in. Sometimes

they do; sometimes they don't. But there's no thinking about what -- the exclusions are an effort to think about what could happen to these patients downstream that are costs that we would want to exclude.

But the risk adjustment model, and as I said, I'm not sure those are inclusive enough, but the risk adjustment model doesn't sort of always capture all the things that are unique to the patient's circumstances that brought them into rehab or brought them into a SNF that might affect downstream costs and utilization.

And it's this lookback, and it sort of ignores -- there's information of the fact this patient went into a SNF. And patients go into SNFs for different reasons. So it's not always clear to me that they have fully captured the circumstances that brought the patient into care that may affect downstream costs. That's my concern, and it's a concern that's driven by the fact that the models are being standardized, even if the coefficients are different. The models

are being standardized.

And the question is should there be more tailoring. And I don't have an answer to that. I raised it because I'm torn and uncertain about this.

MEMBER GLANCE: Isn't there something about the way the CMS HCC models are created that the categories are hierarchical categories? And so they do look at the full set of ICD codes, both inpatient and outpatient, and they group them in such a way that the ones that are most likely associated with higher costs are given more importance?

MEMBER NEEDLEMAN: Yes, but that's exactly right, and that's fairly general. So now we've had a patient who has had a joint replacement and wound up in rehab or wound up in a skilled nursing facility which is providing rehab services. But if the other things that are in that hierarchy are not directly related, you know, in some sense, they get classified out because of something other than the joint

replaced. The fact that the condition that created the incident event was joint replacement sort of gets lost.

MEMBER GLANCE: Okay. Thanks.

CHAIR CELLA: Okay. So this is Dave C. again. So this issue on risk adjustment as well as its exclusions and other topics that have been brought up by the standing committee, will be placed with the subcommittee that anyone can volunteer to be considered for, and we'll do that through NQF. And Ashlie will set that up after the meeting, presumably.

We have hit the time where we're no longer ahead of schedule, but we're still on schedule. And it's time to go to public comment, and then we'll take a break.

I'm not sure how to open up for public comment, but maybe the operator is listening.

OPERATOR: And if you would like to ask a question or make a comment, simply press star then the number 1 on your telephone keypad.

MS. WILBON: Thank you. Those who

would like to make a comment can also enter your comment into the chat box, and we'll make sure it is read or responded to by the Methods Panel. Thank you.

CHAIR CELLA: Okay. So, star-1 or type a note into the chat box, lower left of your screen.

MS. WILBON: And if you're on the webinar, you can raise your hand, and we'll acknowledge you that way, too. At least we'll know that you have a question.

OPERATOR: And again, in order to ask an audio question, simply press star then the number 1 on your telephone keypad.

And at this time, there are no audio questions.

CHAIR CELLA: Thank you. I don't see anything in the chat room. I don't see any hands raised, either.

MS. WILBON: Okay. Great. Thanks, Dave Cella, for leading us through that discussion. Again, very interesting and

thoughtful, and we're looking forward to getting some folks together to write some papers and dive a little bit deeper into these topics and share some guidance going forward.

So we are due for our second break. It is another hour, and we will reconvene at 3:30 p.m. Eastern Time to conclude our meeting with discussion of two measures.

Measure 3559, hospital-level, risk-standardized improvement rate in inpatient-reported outcomes following elective primary total hip and/or total knee arthroplasty, by Yale CORE/CMS.

And the other measure is 0715, standardized adverse event ratio for children less than 18 years of age undergoing cardiac catheterization, by Boston Children's Hospital.

We expect that the developers will be joining us at that time, and we'll come back to discuss those two measures until the end. And that will conclude our day one.

So barring any questions at this

point, we will go ahead and break and reconvene at 3:30.

CHAIR CELLA: Ashlie, this is Dave C. again. Just a note that Colette Pittson (phonetic) has typed in a comment. It doesn't require a response, but just to let people know there is a comment that has been added.

And, Dave N., can you just hang on for a minute?

MEMBER KAPLAN: This is Sherrie. Can I ask one quick question?

CHAIR CELLA: Sure.

MEMBER KAPLAN: Given the difficulties in redialing in, is it better to hang up or stay on the line?

CHAIR CELLA: You can stay on the line.

MEMBER KAPLAN: Okay. Thanks.

CHAIR CELLA: You can stay on.

MS. WILBON: Just if you stay on, just make sure that your phone is muted, please.

(Laughter.)

MEMBER KAPLAN: Yes. Got it.

MS. WILBON: Yes.

MEMBER KAPLAN: Okay. Thank you.

MS. WILBON: Yes, you're welcome to stay on and stay on the webinar as well. Thank you.

CHAIR CELLA: Okay. Bye, everyone.

(Whereupon, the above-entitled matter went off the record at 2:29 p.m. and resumed at 3:30 p.m.)

MS. WILBON: Hi, everyone, it's Ashlie Wilbon from NQF again. Welcome back to our third and final session of the day. This session will be focused on reevaluating two measures by Subgroup 1 that were submitted this cycle. And we're going to dive in here real shortly.

We are working through a couple of technical for folks trying to get on the phone, so we'll continue to do that. We do have a little bit of slides to go through, and we're going to do a voting test as well for those of you in Subgroup 1 who will be voting. So a

couple of administrative things before we dive in.

I also just wanted to check in to see if the developers were on the phone from Yale CORE and CMS? If you want to speak, you will need to hit star one. And you can also chat us through the chat boxes. Please let us know if you've having any trouble getting through on the audio.

I don't hear anyone. I see Lisa Suter on the webinar. Lisa, can you raise your hand if you see me -- I mean if you hear me, sorry. Oh, I see a note from Victoria, you guys are on. Are you dialed into the phone number?

PARTICIPANT: Some people have been joined this way.

MS. WILBON: Can you just chat us and let us know if you're dialed into the phone number. So you won't be able to speak if you're just dialed in through that.

So we'll go ahead and get started, and we'll have our teams work with them, the

background, to make sure we can hear them and they'll be able to participate in the discussion. Just a couple of notes, as I mentioned. The process we'll be using for measure discussion will be the same as we used last cycle, so just a quick refresher.

We'll have an NQF staff person who will introduce the measure, myself or Sam will be doing that today. And we'll have then the lead discussants to summarize some of the key concerns that were outlined in the discussion guide as well as the PAs that were submitted by the subgroup. And then we will invite other subgroup members to provide comments.

We'll then open it up for developers to give a couple minutes of initial response. At that time, we'll ask for the developers to provide a summary of their response, which has been included in the discussion guide in Appendix B. And then we will open it up to the full panel for discussion.

So again, all members can -- all

members of this panel can discuss the measure, but we'll only be collecting the final vote from the subgroup. Any questions about that? And we will, as staff, be noting any folks who will need to be recused from the discussion, and that we have everyone on the line and staff can establish quorum before we vote and discuss.

CHAIR NERENZ: And Ashlie, just one additional small point, Dave Nerenz here. I'll take care of watching the hands up and calling out people. I know Sherrie and Joe are leading the discussion, but they may not be able to see that function, so I'll take care of that for them.

MS. WILBON: Okay. Thank you for that. The voting is going to be done via our Poll Everywhere software, which Hannah just sent an email to all of the Subgroup 1 members to make sure the instructions were at the top of your email. So if you all could kind of direct yourselves to your email and make sure you find that.

There's a link in there for you to click and make sure that you can get into the voting software. You may not see anything right now, but we'll be running a test here shortly to make sure that everyone can vote and that the software is working, and it's ready for the time in the discussion that we'll be ready to vote.

Just also noting that we will over the next -- the rest of today and the remainder of the meeting tomorrow, our discussion today will be focused on measures that were consensus-not-reached, and measures where they did not pass but have submitted -- the developers have submitted additional information for the Methods Panel to consider prior to submitting a final vote on the measure.

So we will not be discussing all measures. Again, just seven measures that will be discussed as part of this meeting. The remainder of the measures that are not being discussed have all passed based on the initial preliminary analysis and will not be -- and have

not been pulled for discussion and will not be re-adjudicated over this call.

Those measures, by virtue of not being pulled for discussion by the Methods Panel members, have passed by consent calendar by the Methods Panel, and the vote submitted for the preliminary analysis will stand as the final vote for reliability and validity for the measures. That is until the standing committee evaluation.

And just a note here that we will be considering transitioning to a full panel vote next cycle. If you recall, last cycle we did the shadow vote and found that there was actually quite a bit of consistency among subgroup members and those members who weren't assigned to the measure the subgroup members.

So we're looking into that, and it's a process change that we're looking into potentially making in the future. But for now, it will just be the subgroup. So I would definitely encourage members of the panel who are not on the subgroup to try tune into the

discussion.

Again, some of the issues are overarching. And even though you were not a part of the subgroup, we would still like to hear any input that you have on the measure based on your review of the discussion guide and listening to the discussion.

So with that, I'm going to hand it over to Hannah to do a voting test for subgroup members, and we'll make -- we'll move into the measure. I'm going to hand it over to Sam after that to get started. And in the meantime, we'll work on making sure the developers are able speak.

MS. INGBER: Okay. Thank you, Ashlie. This is Hannah, can everyone hear me?

CHAIR CELLA: We can.

MS. INGBER: Wonderful. Okay, so we'll be doing a voting test first. I'll be opening the test vote now. If you have the link open on your desktop or your phone, you should see the question: during your morning commute,

what is your preferred method of transportation?

Okay, so I see six votes. So before we move forward, I'd like to confirm that -- which Subgroup 1 members are on the line. I'm going to read through your names, if you could just reply and let us know if you're here.

Daniel?

MEMBER DEUTSCHER: Yes, I'm here.

MS. INGBER: Dave Cella.

CHAIR CELLA: Here.

MS. INGBER: Matt Austin.

MEMBER AUSTIN: Here.

MS. INGBER: John Bott.

MEMBER BOTT: Here.

MS. INGBER: Joe Hyder. Okay, Joe's not in. Patrick Romano?

DR ROMANO: Here.

MS. INGBER: Thank you. Sherrie?

MEMBER KAPLAN: Here.

MS. INGBER: And Terry.

MEMBER WARHOLAK: Here.

MS. INGBER: Michael Stoto? Oh, Mike

is not available at this time actually. And Larry.

MEMBER GLANCE: Here.

MS. INGBER: Okay. Wonderful. Thank you, everyone. So we should be getting eight-something.

CHAIR CELLA: This is Dave Cella, I just have a quick question. So when it says clear last response, if you click that, does that mean it registered the response, or it clears it and didn't register?

MS. INGBER: That means it clears it and it didn't register. So don't clear your last response.

CHAIR CELLA: So when you vote it just -- the vote just stays up on the screen. It doesn't clear.

MS. INGBER: That's right.

CHAIR CELLA: Is that right?

MS. INGBER: It will only change once I shift it on the back end.

CHAIR CELLA: Okay. Thank you.

MS. INGBER: I'm registering nine votes. But is there -- did Joe or Mike join us? I'm not hearing any response.

CHAIR CELLA: This is Chicago. Maybe somebody voted twice.

MS. WILBON: Hannah, maybe try doing another test, then --

MS. INGBER: Okay. Okay, yes. I'm going to clear the responses, and we'll ask you to fill in your vote again.

Hi, sorry for that delay. I'm still registering nine votes. The team is just going to take a minute just to adjudicate this.

DR. ROMANO: This is Patrick. I might be the guilty party. So if --

MS. INGBER: Okay.

DR. ROMANO: It doesn't prevent me from -- in other words I'm on my phone just as a backup, in addition to being on the laptop. But it didn't seem to block the second vote. So is that -- that's the way the system works, it allows two votes?

MS. INGBER: Yes. That is possible.

DR. ROMANO: Oh, okay, I'll do that again.

MS. INGBER: Okay, maybe we should try one more time. Oh, go ahead. Oh yeah, that is confusing. Sorry about that. Okay, we'll try one more time, and we'll get -- we'll hope for eight. Thank you for your cooperation.

MS. BALESTRACCI: Hi, this is Katie Balestracci. Can you hear me?

MS. INGBER: Yes, we can.

MS. BALESTRACCI: Wonderful, I am on. Thank you, hi. I will be the lead discussant for Yale CORE for this measure. Thank you for your patience, I was having trouble being heard.

PARTICIPANT: Thank you for joining us. Apologies for the difficulty.

MS. INGBER: Okay, I'm registering five votes now.

MS. WILBON: Hannah, can you tell who the votes are from? Like if you got a double vote from someone, would you be able to tell who

it was, or?

MS. INGBER: No, unfortunately not. Waiting on one more vote. Okay, we have eight votes now. Thank you so much for your cooperation and your patience with that. I'll hand it over to Sam.

MEMBER SIMON: Excellent. Thank you so much, Hannah. I'm glad that we've got our testing up and going, and hello friends and colleagues. It's been a little while since I've had the pleasure to be with the Scientific Methods Panel.

We've -- Ashlie has been riding solo for a bit here, and it's undoubtedly to my detriment. I don't get to interact with you nearly as much as I'd like to, so big thanks to Ashlie for inviting me to join you today to introduce this measure.

Just a quick reminder in order of operations. I'm going to introduce the measure. Lead discussants will then summarize the key concerns. We'll invite the subgroup members to

comment, and then we'll hand it over to the developer to do two to three minutes of initial response. And then we'll open it up to the full panel.

Once we've gotten through that, we'll of course move to voting. And then we'll move on to our next measure.

And just to let, to remind everyone where we're starting from, this is Measure NQF 3559, hospital-level risk standardized improvement rate in patient-reported outcomes following elective primary total hip and/or total knee arthroplasty. So the initial results of this is that the measure passed under liability, but consensus was not reached on validity.

So when we hand it over to our lead discussants, Sherrie and Joe, we'll ask them to focus primarily on the validity concerns that were expressed by the panels who reviewed this measure.

Now, as was mentioned, the developer for this measure is Yale CORE. And you'll find

inside of your discussion guide a summary of the issues that were causes for concern amongst the panelists.

Now, big thanks to Yale CORE for putting together a very robust response to the concerns that were identified by the panel. That was attached to your meeting materials as well, and I hope that you have all reviewed that. But I imagine that we'll be going through that in some detail.

Just as a couple of highlights about the measure, the testing that was conducted -- excuse me just one second. The testing that was conducted for reliability was both at the measure score and the data-element level. So looking at the test-retest reliability of the hip dysfunction and osteoarthritis outcome score for joint replacement or HOOS, JR., as well as the knee injury and osteoarthritis outcome score for joint replacement, KOOS, JR. ,as it were.

They also did assess consistency, and those produced some pretty good results overall,

hence the panel not identifying too many problems with liability. Signal to noise analysis was conducted at the score level.

I won't go to the main concerns associated with reliability as those were the past reliability, and those are largely addressed by the developer. But we can wade into those during discussion if the panel wishes to do so.

Now, on validity, as I mentioned, consensus was not reached. This was, as is appropriate for all PRO-PMs, this measure was tested at both the data element and score levels for validity. Looking at responsiveness, external validity, and floor and ceiling effects for both the HOOS, JR., and KOOS, JR. instruments.

I won't go through the results in too much detail because I think our lead discussants will likely want to get into that themselves. So at this point, I'll hand it over to Sherrie and Joe to lead our discussion.

MEMBER KAPLAN: Who wants to go first?

MEMBER HYDER: You can arm wrestle on that one, Sherrie. As you chimed in, why don't you go ahead.

MEMBER KAPLAN: Oh, dear. Okay, so one of the first questions I had for the measure developer and for NQF was it -- this is created as a composite score, and I am always concerned that I'm getting NQF's definitional things incorrect. Because when you're creating a composite across multiple items that are still the same construct, does NQF count that as a composite? In this case, knee and hip function.

MS. WILBON: Yeah, this is Ashlie. I think that's a good question. It's actually a question that we have that I think Dave Cella brought to our attention and is actually at the bottom of the slide about whether or not this measure qualifies as a composite.

I think, and Sam, I'm going to ask for your help a little bit on this, because I think we have had this question come up before with programs that have multiple items. And I -- do

we also sometimes classify them as both or can it be both? Or, I'm going to lean on you a little bit for this at this point, Sam.

MEMBER SIMON: That's really tricky question, I wish I had a better answer for you. And I can't think of examples of where we've had multiple items that could potentially roll up into a single score that I've used -- I've used, as in this case, to effort scoring instrument. So I unfortunately don't have a good answer.

MEMBER KAPLAN: Well, here's why it's relevant. Because, and it would be a little, it would -- for people like me, it would be weird to have data-element reliability.

For example, in these multi-item things that are supposed to assess an underlying construct like knee function or hip function, so you wouldn't go after, that would -- you wouldn't go after data element and reliability anyway. And so you're only doing score-level reliability.

And then this, there's the patient and the hospital or whatever facility level,

reliability questions too. And that was of concern when I was thinking about the reliability piece of this. You know, what happened to the across patient -- across items within patient part of the -- in the denominator.

And so when you're thinking those things through, now that, because we haven't had that discussion yet and it would be unfair to apply those new, you know, criteria around those things. But if we're creating composites, that piece of the error term, when you're aggregating it up to the next unit, it becomes very relevant for the error term and how you're comparing things to things like validity variables.

So that's probably, if we can't answer for this round, then we have to just ignore it. But that's something probably we should put in a parking lot and come back to in future discussions.

So having said that, there is concern raised, and I'm reading this from the notes from others and the compiled notes on concerns raised

by the S&P on validity. One is attribution, and it's relevant to this business about well, where is the error variance and where is the real two-score variation.

Attributing joint function, changes differences in joint functions at a hospital versus rehab services, verus something that can happen in the interval as long as nine months to a year following surgery.

So which piece of the variance belongs to the hospital in terms of attribution, versus whatever else they got after the surgery with that long of a time interval. So that's one concern that was raised. Shall I go on, or do you want -- shall I just blast through these and then have, ask the developer to respond, or do you want to do them one by one?

MS. WILBON: Sure, Sherrie, this is Ashlie. Go ahead and we'll see if Joe has anything to add, and then we can open it up to the developer.

MEMBER KAPLAN: Okay, there's a

concern raised that suggests the exclusion of staged procedures made up to 43% of the procedures. And the measure should include, from unstaged procedures. I'm not exactly clear what that refers to, but, and it wasn't, I didn't bring that one up. But that was another concern raised.

The exclusion analysis noted that data weren't provided in the same way that there was -- a concept diagram would have helped on how much the exclusions affected the performance scores. There were questions raised about the 25-case volume threshold and what that based on, and what happens when a facility falls below the 25-case recommendation. Were they excluded or did you do some other kind of analysis on those data, and did you examine the inclusion/exclusion of those 20, you know, the smaller volume hospitals.

Another concern that came up was about risk adjustment. The model was including hospitals wasn't used for reliability, validity, and missing data, for example. And I'm not

reading this one right. Hospitals with low caseloads were not recommended from the measure. Did the developers -- oh, did the developers do a sensitivity analysis, which is related to the issue I just read, that tests the impact of including hospitals from risk adjustment, the risk adjustment development sample.

And then another concern was meaningful differences. There was some question about how much of the variation could be -- establish meaningful differences between hospitals in the top quartile for example.

And finally, there's an issue about missing data. Two people raised concerns about missing data, and that only the complete data were analyzed without accounting for what's likely very, fairly extensive missingness. And whether that's actually missing at random or missing not at random. And so that's kind of the gist.

And I got the -- it says measure developer response in the piece I'm looking, but

I didn't get the robust response that Sam was talking about, so I don't know where that lives. But apologies for not having read that in advance.

One other question, one other comment is the data-element validity presented the first of studies in the literature at the patient level. But again, I'm not concerned about that because this is supposed to be s composite measure.

The end.

MEMBER SIMON: Ok, Joe, did you have anything else that you would like to add?

MS. WILBON: Hi, Sam, I think we're trying to figure out if Joe was on or not. I thought I heard him, but Hannah was saying she thought he was not on. So if that's the case, then we can open it up to other subgroup members.

MEMBER SIMON: Okay now, so let's go ahead and open it up.

CHAIR CELLA: Okay, well, this is Dave Cella. Maybe I'll start then. I'm one of the

subgroup members. And Sherrie really focused on the measure, measure validity and wasn't concerned about data element. But I'm actually quite concerned about the data element, and maybe I can explain why.

So, and so I'm wondering if I should talk in the weeds or up some 30,000 feet. Now, let me start at 30,000 feet and then get into a little bit of the weeds.

At the data-element level, a little background for those of you who don't know the HOOS and KOOS very well. The HOOS and KOOS, JR, were developed kind of off the grid of the developers of the HOOS and the KOOS as (telephonic interference). They combined pain and function into one measure, so, and they do that I guess to have one score and a short measure.

And by doing that, given that this is a performance measure that really relies upon patient self-report, by doing that, they conflate pain and function. And in some -- in many cases,

particularly with knees, the recovery is different for pain as it is for function.

And although they used a rush measurement technique to create (telephonic interference) that would hold together well at least in the sample they tested, that's not necessarily how they'll perform later.

So there'll be a lot of error introduced. And it's a short measure with a pretty good sized ceiling effect, so it's not clear that it'll work very well for people that come in at the relatively high function or low pain level. Again, mixing the two.

So one of the things that the HOOS and KOOS developers did in collaboration with the late Barbara Gandek was they went ahead and developed short versions of the HOOS and the KOOS that keep pain and function separate. But those short versions were developed and published after CMS had allowed the JR to qualify for bundle payment. So JR kind of got in the hopper.

But there are these, I would think

from a PRO person's perspective, better measures that are available, and actually in the fourth data set, that the developers could use to allow pain and function to remain separately measured with still short measures. The reason they went with the very short measure was because the TEP recommended it.

But there's, and another, you know, maybe less important issue but I'll mention it is that the JR versions, the ones that are being proposed, are really only using orthopedics and only in the subset of orthopedics. Not using rheumatology, for example, where there are clinical quality measures that rely on the same issues of pain and function in the hips and knees for osteoarthritis.

So I think there's a real strong case to be made for reconsidering using these 12-item versions that are, you know, they take two minutes per person as opposed to one minute per person, so it's not that much more of a burden. And I'm not sure how much the TEP was included,

was asked about using the 12 versus the six-to-seven item JR version.

But one of the other benefits, and this is getting a little bit into the weeds, and I'll remind everyone about my conflict, that at the beginning of the meeting, I am on the board of a nonprofit, PROMIS Health Organization. But one of the other benefits of having pain and function separately measured at the patient level is you can then allow the orthopedic community flexibility in what they use.

They can use the 12-item version or the full HOOS and KOOS. They could use a PROMIS measure. And many of them have already converted over to PROMIS measures because they don't have that problem with the ceiling effect.

So you know, I think it's a good submission and I came down it on low on validity because I think there's a much more valid way to get information from patients about pain and function, reduce ceiling effects, and leave more assessment available to the community, so that

they can use the 12 or they can use PROMIS, or anything that can be linked. WOMAC, including WOMAC, that could be linked to these pain and physical function measures.

We'll make a happier orthopedic community, make for better registry connectivity, and better understanding down the road.

Otherwise, what's going to happen is that everyone will converge on the JR mixing pain and function together and this opportunity will be lost. So that's why I came in where I came in on this. There are other, you know, minor comments I might chime in with based upon where the discussion goes, but I'll leave it at that.

MEMBER SIMON: Thanks, Dave. I've got Matt, Larry, and Sherrie in that order. And I almost will check with Sherrie as the discussion leader if you want to stay at the spot in the queue or if you have some immediate response.

MEMBER KAPLAN: Well, I kind of -- this is Sherrie, I kind of do have an immediate response, because it seemed like, Dave, you were

raising some separate, some separable issues. One is dimensionality within the hip and knee osteoarthritis outcomes measures.

So there's dimensionality because for example, the knee includes a stiffness dimension, that's a single item, that's its stiffness, plus pain and daily function. And the dimensions are differentially, you know, represented with item content. But are you raising a concern -- I was kind of under the impression that we were evaluating what we're seeing, not what we could see.

And that's a question for NQF, are we evaluating the measure as we're staring at it, or are we evaluating it in a larger context of how, what would you choose if you were, you know, opting for the ideal measure for, of hip and knee function post-intervention and recovery.

Go ahead, Ashlie, it's all right.

MS. WILBON: No, no, no, sorry, I was just going to say we're evaluating the measure as is and as it has been presented to us.

MEMBER KAPLAN: Okay, and then so my question back to Dave Cella is are you raising concerns about multi-dimensionality within the knee survey and whether those have been, are accurate for reflecting in terms of content validity, reflecting all the content. Is that the concern?

Or is it -- because if you shorten some of these measures as we did for the SF-12, for example, you destroy the internal dimensionality, but you still have a, have construct validity at the overall score level. I'm a little bit confused about --

CHAIR CELLA: Yeah, I, my vote was low on validity, and it was four reasons, really, one of which was this missed opportunity. But three reasons that I gave before that are the concerns about the missing data, the proxy reporting issue, and the ceiling effect of the data element.

MEMBER KAPLAN: Okay, great, I just wanted some clarity. Thank you.

CHAIR CELLA: Yeah.

MEMBER SIMON: Okay, Matt.

MEMBER AUSTIN: Yeah, good afternoon, everyone. So I have maybe a couple of questions for maybe NQF staff more than the measure developer. And that's just to sort of interpret maybe two sort of rules or guidelines.

One is for the validity testing, they provided those measures, score validity testing and data-element validity testing. If we're comfortable with one but not with the other, how should we proceed?

And the second question is when we talk about data-element validity testing and needing to look at all critical data elements, does NQF have a definition in what's included and not included as a critical data element?

MS. WILBON: Hi, Matt, this is Ashlie. So given the different types of -- for your first question on the different types of validity testing, I would say, you know, if there is a concern with one kind of, one of the kinds of

validity testing, that you should incorporate that into your evaluation, similar as you would of considering, you know, all the various components of validity, you know, risk adjustment, inclusion, missing data.

All of those components are also kind of subsets or subsections of validity. So we, you know, we don't have any specific guidance on how those should be weighed, but you certainly should take into consideration how all those different components kind of, you know, weigh against each other and whether or not that concern is significant enough to impact your vote, you know, one way or the other.

MEMBER AUSTIN: Okay.

MS. WILBON: So the second question around critical data elements, I was actually looking for the specific language. But it's essentially, I mean, we consider, you know, the critical data elements, those data elements required to compute the measure, so essentially all of them would be critical. But definitely

the numerator data element.

And I'm sorry, I can't quote it off the top of my head, but give me a second to find it and I can find it for you.

We don't have any specific requirements on, you know, if they did five of the six or, you know, how we, you know, we certainly have submissions where, you know, they've done, you know, maybe five of the ten data elements and those ten reviewers have found that to be sufficient.

I think the word critical is somewhat subjective and our language is not quite definitive on that, although we do provide some examples in our description of how we define critical data elements. And I'll find that language for you in just a second, and I can share that fact with the group.

MEMBER AUSTIN: Okay, that'd be helpful. I just want to make sure I'm being fair in how I'm applying that term, critical. Okay, thank you.

MS. WILBON: Sure, thank you.

MEMBER SIMON: Okay, Larry's up.

MEMBER GLANCE: Hi, everybody. So the caveat here is patient-reported outcome measures are not something that I know an awful lot about. And I'm sure Sherrie will point that out at some point. Hi, Sherrie.

But a couple thoughts. One, when I looked at this measure, it struck me that -- I think this is a CMS measure, correct?

MS. WILBON: Yes, I believe Yale was contacted by CMS.

MEMBER GLANCE: Right. So it struck me that if you look at the measures that are currently being used in the comprehensive joint replacement program by CMS, that I think the only outcome measure that they have currently is the risk standardized complication rate.

And I think that when I evaluate measures, a couple things that I look at, and I know this is not typically what we, you know, the typical schema. But I look at things like how

much of a need do we have for a measure like this.

And my understanding of most of the episode-based payment, the bundled care programs that CMS has, as well as the accountable care, the qual metrics that'll be used for ACOs, I don't know of very many, if any, patient-reported outcome measures. So there's a barely a need for this kind of a measure in my mind.

The second thing that I think of when I evaluate a measure is, you know, we talk a lot about cutoff values, criteria that we would like to be able to use. And I think that's kind of hard.

But one of the things I do is I sort of compare our measures when I'm evaluating compared to all the other measures that I've looked at for the last couple years, both on this Methods Panel, as well as the standing committee that I used to serve on.

And so I look at the quality of the measurement. I know that that's sort of a gut-

level thing, but to me that's important because I just sort of put it into context.

So when I looked at this measure again, it occurred to me that there's a need for this type of a measure in the performance metrics that are being used by CMS. Because what we need to look not just at complication -- and costs and process measures, we also need to look at patient reported outcome measures.

And secondly, I looked at the quality of the measure, and I kind of thought it was at least as good if not better than the majority of the measures that I've looked at. And I think the one that consensus not reached was on validity, and I'll sort of limit my specific comments on validity.

So even though this is a composite measure, it boils down to it's an all-or-none measure. So it's a, essentially ends up being a logistic regression model. And they used the standard approach to looking at the validity of the prediction model in a validation data set,

and they found reasonably good discrimination with a P-value of 0.69. Not terrific, but it's not uncommon to see C-stats in that area for models for any kind of, anything other than mortality. Mortality models typically have C-statistics that are better. But other models that are not mortality models, and this is not unusual.

And then the calibration, they looked at calibration using calibration plots, and those seemed pretty good. And they also used another standard approach as well, and the calibration there was good as well in the validation data.

So to me, at least that the score-level validity seemed to be good. And I think we were all agreed that score-level reliability was okay. So that was the reason that I voted to pass this measure. And again, I don't think it really matters whether it's a high or moderate. But in my mind it's sort of a binary decision, and I voted pass for those reasons.

MEMBER SIMON: No hands up. Back to

Sherrie and Joe again.

MEMBER KAPLAN: I always have to be careful that I don't hit the off button while I'm trying to hit the unmute button. So, yeah, I think the big issue I think for me is the missing data. And I think Dave, at least Dave Cella, raised this question, that more than 50% of hospitals with more than 25 eligible patients and about 70% of all hospitals had missing either PRO and/or risk variable data.

And so despite their propensity for analysis where they compared those with and without missing data, there still remains concern about a pretty substantial bias in the data presented. And so that was, that I think is a, remains a big concern of mine.

And then I was also concerned that there's the association with the risk-standardized complication rate was the NQF 1550 was described in the text as a correlation. But the data presented were, figure 1, for example, had a box plot. So, and it wasn't clear the, I

may have missed it, but it wasn't clear the analysis that generated these data was provided in detail.

And so there's considerable variability at the mean and a plot of the association of sort of a pass/fail of each one of these measures at the hospital level would have been really helpful, along with the standard error bar. Even in the 123-hospital sample, if we could just have seen the magnitude of the error there, it would have helped sort of figure out how hospitals compare to each other and how this might be used or not used.

And again, I did miss the response from the measure developer. So if that was addressed in that response, I apologize.

CHAIR CELLA: This is Dave Cella again. I think we're supposed to turn to the developer since there are no more comments. But just to say that my, you know, I share the concern about missing data. I also think there shouldn't be proxy reports used.

But the other thing is the orthopedic community is really getting on board with patient-reported outcome data. So I'm optimistic that in the future there will be better quality data collection, because at present there is compared to what, to the data that were analyzed for this submission.

One of the problems is that, you know, what's being used isn't just this one particular set of questions. And this one particular set of questions is not going to be easily be mapped to others. Whereas a different choice that could be done with the same data that were used here in the FORCE-TJR registry could be linked to other things that are being mentored now.

So it's, you know, maybe a somewhat out-of-order suggestion to the developers that a change in the actual element could bode very well for a good participation rate in the orthopedic community and we move that missing data problem downstream. That otherwise I think we'll still be there because people are committed to other

measurements.

MS. WILBON: Thanks, Dave. Yes, let's open it up to the developer. And I did want to just point out, because I think perhaps there was a question, the developer's response is included in Appendix B of the discussion guide. It starts on page 38.

We would ask the developers, if you could, you know, summarize your responses in Appendix D. I think many of the subgroup members reviewed it, but I think some may have not had an opportunity. So if you could do that, that would be very helpful.

MEMBER KAPLAN: Actually, Ashlie, could I also ask that the developer speak to the issue of attribution to the hospital, given the time window following surgery is a fairly broad nine to twelve months, is quite wide. And that could have a lot to do with rehab and other things. Can they also address that, thanks.

MS. WILBON: Sure, thank you.

MS. BALESTRACCI: Hello, this is Katie

Balestracci from Yale CORE, can you hear me?

MS. WILBON: We can, thank you.

MS. BALESTRACCI: Wonderful. So thank you for this opportunity to speak. I would like to start by just giving the context with which this measure was developed, specifically in terms of stakeholder input.

This measure was developed, testing results were presented to, and a great deal of input was provided by, our technical expert panel. We had a technical working group that has worked with us over the life of the development of this measure. There were stakeholders versus public comments.

During the creation of the CJR model under CMMI, which was the comprehensive care for joint replacement model, which was the mechanism by which these patient-reported outcome data were collected and submitted, required and developed a great deal of input from our TEP, but also from the hip and knee societies in terms of data to be collected, data elements to be included for risk

adjustment, and the PROM survey instruments used.

I noted earlier a question by one of the panel members about the choice of PROMs. There was a great deal of discussion early on from the TEP about different PROMs. There was an evaluation of -- including the WOMAC, including HOOS and KOOS the full version, etc. So these decisions, the decisions around PROMs, the decisions around the data elements chosen went through a particularly rigorous stakeholder examination evaluation. And we received a great deal of input.

Likewise, there is for this measure a particularly important feedback and input from patients. Not only did we have patients on our technical expert panel, but we engaged a patient working group that would allow us to get input specifically from a larger number of individuals who have undergone either a total hip or a total knee arthroplasty, and in several cases multiple such procedures. So I did want to put this in that context.

As I listened I've heard a number of different concerns around the validity of this measure. I will say that each one that was mentioned in the last half-hour was addressed in the 12-page response we provided to the panel. But I'm happy to try to quickly address them here. And then ask that, you know, if someone needs even more detail, I'm happy to provide it.

I heard concern about attribution of this measure. This is a hospital-level measure because we are hoping with this measure to attain the following goal: To really capture the full spectrum of care and to incentivize collaboration and shared responsibility.

I think we note that a surgeon can do a terrific job, but if there is not a coordination of care around other providers within the hospital, around discharge planning, etc., then a patient may not receive the positive outcomes that they are expecting and that their surgeon is expecting.

So we are really expecting that this,

as a hospital measure, will allow patients to then receive a high quality of care across that spectrum. It is a follow-up. The follow-up post-assessment is a nine to twelve month post-assessment, because we do want to make sure that patient has had the time for a full recovery, whether it's a knee or a hip replacement.

And our technical expert panel and other consultant experts indicated to us that they thought that follow-up would sufficiently cover that.

I heard a concern about staged procedures. I will tell you that we are unclear where the 43% number came from in terms of the panel. It is our experience in our data that the number of patients who may have these type of procedures, staged procedures within a year of each other, is about seven percent. We noticed also globally there are some studies coming out of Sweden that mirror this exact percentage.

I will say that in our data, 4.17% of the patients with complete PRO and risk-variable

data were removed because of the staged procedures. And we did not do a sensitivity analysis for this reason.

Our very concern about the staged procedure within a second procedure within a year of the first is that we don't know how to interpret that outcome because of the overlap of the recovery period.

I heard a concern about a 25K volume threshold. This threshold was chosen because we expect that in public reporting, such a threshold would be applied, and certainly it is applied in many of the claims-based measure with which this measure is intentionally harmonized. So this is not an exclusion, this is in fact a reporting and expected or recommended reporting threshold.

I will say that from our data, while 33% of hospitals conducted fewer than 25 elective primary hip or knee procedures, the number of procedures that this represents for the overall number of hip or knee procedures that occurred across the country in hospitals during that

measurement period is 3.14%.

So while hospitals with fewer than 25 procedures appears to be around 33%, it was only three percent of procedures themselves that then did not get included in the measure.

I heard a concern about the choice of HOOS, JR, and KOOS, JR, and I will simply reiterate what I said as I began speaking, which is that both of these PROMs went through a rigorous examination by our technical expert panel, as well as a public comment period when the CJR model was being proposed. And they continued to be the measure that was supported through the development and final testing of this measure.

I heard a concern about missing data. In this measure, one of the things that we recognize as developers of a patient-reported outcome-based performance measure is that unlike claims-based measures, the data need to be taken directly from or gathered directly from patients, so that there is a voluntary nature to this.

We have proposed in our responses, and certainly as we continue to consider PRO-PMs, there is certainly ways to improve the response rate. And we believe that the CJR model has really provided a proof of concept of how this can be approached. We do have for this measure, as you note, an approach to response bias.

We recognize that as the collection of PRO-PMs increase as hospitals get better about collecting them, as providers get better about collecting them, you know, likely that will never be or hard to get to 100%. So we do believe that that has to be something we pay attention to, and so we did do a statistical approach to address potential response bias.

You will note in our results that we did not find a large difference in measure results when we applied this response-biased approach statistically, but we believe it is an important consideration and recommend that that be part of the measure.

I will also note that we did, in the

response-biased approach, identify and deal differently with both cases in which no PRO data were provide, and cases in which incomplete data were provided. So addressed both missing and non-response, which are of course two different issues.

Let me just address what I think are the two other things that I heard. There is a concern about proxy reporting. We expect in an elective procedure, such as this elective primary hip or knee replacement surgery, that proxy responses will be quite low. They were certainly low in the data that we used for the development of this measure.

However, because of the nature of volunteer reporting, we did, and this is part of the requirements around the CJR final rule and data that will be collected, we did allow for proxy responses because it did provide us with additional information for patients undergoing this procedure.

And lastly, let me address critical

care elements. We also provided some detail in our response when the Scientific Methods Panel noted this concern. We have, as you saw, provided the data-element reliability and validity information on the HOOS JR and the KOOS JR.

The claims data that are used both to identify procedures in the denominator and also are used for some of the risk variable co-morbidities are those which are regularly audited by CMS.

We have other claims-based measures in which we have done medical chart review in order to validate this type of data. So there have been -- there are many ways in which we can have confidence in these claims data.

With reference to the data elements that are risk variables in the model that comes from CJR hospitals through the collection of patient-reported outcome data, those particular risk variables were vetted by a thorough literature review by our technical expert panel.

We did a medical chart review that looked at the reliability and the feasibility of these variables in hospital data collection. So given those inputs as well as continued support from our TEP clinical experts and our patients, we feel confident that those variables are in fact valuable choices here.

I am going to ask if Lisa Suter has anything to add before I conclude my comment.

MS. SUTER: Can people hear me?

CHAIR CELLA: A little faint but yes. Okay, I will turn up my volume. Is that better?

CHAIR CELLA: Better, yes.

MS. SUTER: Great, thank you. First of all, I want to thank you all for the opportunity to comment. Katie addressed very robustly the issues raised. I wanted to provide just a little bit more historical context about the concerns with -- I'm getting a lot of feedback and background noise.

But I just wanted to let people know that this measure, the CMS embarked on this

measure many years ago, even before the comprehensive care for joint replacement model.

It actually, the initial technical expert panel didn't have a short form version and made a decision to go with the HOOS and KOOS full form version. The technical expert panel recommended those, given the other limitations and proprietary nature of other surveys at the time.

Public comments and orthopedic stakeholder feedback actually were the ones that identified the short form versions that were developed at the hospital for specific -- the Hospital for Special Surgery, where they actually created the short form -- the surveys with the actual joint replacement patients undergoing joint replacement, as opposed to the HOOS and KOOS, which were developed in knee osteoarthritis patients broadly.

And the patients were deeply involved in the development of those short forms. Those short form surveys were brought back to CORE and

CMS with a request that we consider shorter forms to minimize burden. And through a tremendous amount of stakeholder input, as Katie already alluded to, the decision was to move forward with those short form versions.

I want to say in addition to a patient work group and patient technical expert panel members and two public comments, we have also had multiple orthopedic societies, including the AAOS and American Academy for Hip and Knee Surgeons and the Hip Society and the Knee Society, comment on this measure, provide feedback on this measure, have representatives engaged in this measure.

And each of those organizations, you know, supported the measure moving forward with shorter forms, acknowledging that there are some limitations. But you know, there were individual orthopedic surgeons with communication, with the leadership of those organizations involved in providing advice to us as developers.

I'm not saying that those societies

endorsed these measures. I would not, you know, be so bold as to say that. But they have engaged in the development and provided us feedback throughout.

So we are very appreciative of all the engagement that the orthopedic community has given us throughout the development of this measure, including a summit where the orthopedic community came together and recommended the limited number of high value risk variables to collect prospectively as part of the joint replacement model.

So I just, it's hard to reflect how much went into building this measure, and you know, I'm sure all measures can be improved, and we are very happy to hear the committee's feedback about how to improve this measure and we want to make it better.

But recognizing that this measure has had a, you know, almost a ten-year arc from its inception through having data collected prospectively through a voluntary model that

incentivized hospitals to collect the data, not require them to participate in, you know, in a proprietary registry or other, you know, treatment-requiring process.

But to submit the data in, you know, in many different ways that allowed them that flexibility has been a -- it's been a privilege to be a part of that. And I just wanted to ask people to have a little additional context to what went into the development of this measure. Thank you.

CHAIR CELLA: We appreciate that. Just a very important time check here. We're 25 minutes to the hour, we have one more measure yet to get in before the scheduled adjournment. Sherrie, do you see a path to getting your group to voting pretty quickly here?

MEMBER KAPLAN: I think so. I have two remaining, and it's just a remaining concern, because I have now had a chance to go over the response by the developers and listen to their presentation. I have two remaining concerns.

One is attribution, because even though the TEP is a very useful thing to have for face validity terms, it would have helped a lot to have a within versus between facility variance -- inter-class relation coefficient.

That would have helped a lot with the attribution issue, because I think a year out, a lot of things have happened that can contribute to the patient's outcome that may or may not have anything to do with the hospital, including social support and other things and so on.

So, and I do remain concerned about missing data, because HCAHPS has been around for a long time and now patients are getting more and more surveys. So you know, using these things on the ground has to be held in context with a 15-30% response rate that HCAHPS remain stuck at.

So those are just two remaining concerns, and I don't have anything further to add. I don't know if Dave Cella or Joseph or anybody else has anything remaining.

CHAIR CELLA: Well, if I could, this

is Dave, I just want to say I have a lot of respect for the Yale group, and Lisa and Katie in particular. And I've worked with them in other contexts.

And so, and I understand probably more than most what it's like to have a seven- or a ten-year arc of development work. And so, you know, I said what I said, you know, with that knowledge and not to take it lightly.

I wouldn't have said anything if it was just a matter of is it good enough. Yeah, it's good enough, it's probably going to go fine. But the downstream implications of not doing what I think can be done with existing data are significant, and it's unfortunate. That was really, you know. And I think the committee just has to decide if that's what it wants to vote on or the missing data issue or not.

I just, I think this can be fixed in a way that keeps it much more open for the orthopedic community. They're going to be a much happier group.

And maybe I'll put it to you a different way. Is there, can you see any way two, three years down the road, to be able to say, okay, you don't have to administer the HOOS KOOS, JR. You could administer the 12-item short forms and the, or the PROMIS and approximate the same benefit. And you know, let's link those up.

I don't have a lot of optimism that that can work or that it would ever be done. And that'll, you know, it just changes the landscape forever. Although I do, you know, listening to Lisa's description, I can see what happened all along the way and it made sense.

But it's a little bit now like, you know, like I'm saying there's a better drug. And are we going to go ahead and say this is the drug that's going to go in the formulary, or are we going to put the better drug in the formulary because it's right there and it can be shipped.

So it's, you know, I don't know if that's in or out of scope, frankly, and that's part of why I wanted to have this discussion.

This was actually going to be one of the general topics for discussion, but we decided since it was only related to this particular submission that we would have it here in the submission.

MS. SUTER: So this is Lisa Suter again. I'll just acknowledge the -- and I don't know if CMS is on the line. I could not possibly speak for, you know, what is going to happen with this measure two or three years down the road. But all CMS's measures are maintained with contractors that do annual review and update the measure that reflect the, you know, science that is out there.

And I think this, you know, I would assume that measure would go into reevaluation on a regular basis as well. I don't know how, you know, it's a lot different to make a code update, you know, when the ICD-10 codes are expanded or changed than it would be to change an entire patient-reported outcome.

CHAIR CELLA: Yeah. The only example I have, Lisa, is the PHQ-9, and I can tell you

there it has not been possible to modify. The use of the PHQ-9 is the only measure that qualifies for that particular quality measure, even going through maintenance cycles.

It's part of why I am -- anyway. I think, let's be optimistic that CMS will look at differently than the developer of the PHQ-9 measures, but I don't know.

MS. WILBON: Hi, this is Ashlie. I'm wondering does Subgroup 1 feel ready to vote? I just want to be respectful of the time, and we have another developer group on the line.

I'd like to make sure we have -- we do due diligence for that measure. If there are any other kind of quick, lingering issues we could mention, or if the subgroup feels ready to vote based on the additional information submitted the developer in the discussion today.

MS. BALESTRACCI: Ashlie, I'm so sorry, this is Katie Balestracci from Yale CORE. If I can just note one thing in response to the comment we just heard. There was a concern about

ICD testing or the lack of it. And in fact, the measure score is based on an ICD which about 0.25, my analysts are telling me. So we do want to note that in fact we have that data.

MS. WILBON: Thank you. Okay, Subgroup 1, can we call a vote?

MEMBER KAPLAN: I think so.

MS. INGBER: Okay, wonderful, thank you, everyone. Voting is now open on Measure 3559, hospital-level risk standardized improvement rate in patient-reported outcomes following elective primary THA and TKA. Reminder, this is a rating for validity.

Okay, I will now share my screen to show the results. Okay, as you can see we had five ratings for moderate and three ratings for low. This means the measure passes on validity.

MS. WILBON: Thank you, Hannah. So we are running behind. We had scheduled to review another measure during this time, 0715, standardized adverse event ratio for children younger than 18 of age undergoing cardiac

catheterization from Boston Children's Hospital.

But I know they have holding on the line, and I want to be respectful of folks' time and make sure that we give them an opportunity to review, an adequate opportunity to review and present their measure. Is it Lisa, are you there? Can you hear us?

MS. SUTER: Can you hear me?

MS. WILBON: Yes, yes.

MS. SUTER: Okay, good.

MS. WILBON: Can I just ask, are you guys, I know this is a tough position for us all to be in, and I apologize for running over, would you guys be able to join us tomorrow for a discussion of your measure, or is today, are you locked in for today?

MS. BERGERSEN: We're happy to do whatever works for the committee.

MS. WILBON: Okay. I'd like to, let us take some time as a team and with the co-chairs to look at the agenda for tomorrow, and we will get back to you this evening. And we're

going to add you to the agenda for tomorrow.

I just want to make sure there is enough opportunity for the committee to consider your measure, as well as for you to give, you know, a response and engage in some discussion, while also being respectful of folks' time and ending our day today at five.

So I really appreciate your flexibility, and we'll be in touch very shortly after the close of the call to work out a time that you guys can join us tomorrow.

DR. ROMANO: Ashlie, this is Dr. Romano. I guess there isn't any way to go past five? Because pushing to early tomorrow morning would be 6:00 a.m. on the West Coast.

MS. WILBON: Right. We did actually move Subgroup 1's measures to the afternoon for that reason. So we'll kind of rejigger things around a bit. We have developers who have some kind of restrictions on what time they can attend. And so we'll take a look at things and move things around, and we're definitely keeping

that in mind (telephonic interference).

Dave and Dave, do you have any other comments or thoughts about that for the group before we wrap up for today?

CHAIR NERENZ: No, it's a tough spot. I mean, I feel very badly about asking the developers particularly to adjust schedule, but I also don't want them to feel shortchanged in terms of the opportunity for discussion and for them to talk through some of the responses.

So I think all of us will just have to remember tomorrow to keep the discussion focused, keep the comments as brief as possible, because we'll be squeezing in one more thing beyond what we had originally schedule.

DR. ROMANO: And this is Patrick again. I just, for the benefit of other folks, the developers provided quite a detailed and lengthy response to the comments on pages 49-59 of the discussion guide. So everyone should take a look at that material in advance of our discussion.

CHAIR CELLA: Thanks, this is Dave Cella, nothing to add, that sounds good. Let's try to keep the same time parameters, start and end time parameters tomorrow and we can just be efficient.

MS. WILBON: Okay, the team will take a look at the agenda and we'll touch base with the co-chairs and with the developers this evening to make sure we're on track for tomorrow. And we're at about ten minutes before the top of the hour, and I think we're going to go ahead and open it up for public comment for a few minutes, and then we'll go ahead and adjourn for the day.

And at this time, if you would like to comment, you can enter them into the chat box of the webinar or press star one, although I think we've been having some difficulty with that. But if you're able to raise your hand if you're on the webinar or enter your comment in the chat box, we'll be sure to relay your comment and find a way for you to (telephonic interference).

Okay, I think, hearing none, I think

we'll go ahead and adjourn for the day. I did just want to thank everyone for being such troopers and being engaged for such a marathon webinar for today.

I do think this having been over webinar I don't think that we lost very much by not being in person. So I really appreciate that, and we look forward to more discussion tomorrow and getting through all the measures.

So thanks, everyone, and have a good evening, and we'll talk to you tomorrow morning.

(Whereupon, the above-entitled matter went off the record at 4:51 p.m.)

**E**

**Q**

**S**

**U**

C E R T I F I C A T E

This is to certify that the foregoing transcript

In the matter of: Scientific Methods Panel
Spring 2020 Meeting

Before: NQF

Date: 04-01-20

Place: teleconference

was duly recorded and accurately transcribed under my direction; further, that said transcript is a true and accurate record of the proceedings.

Neal R Gross

---

Court Reporter