National Quality Forum
Scientific Methods Panel Measure Evaluation Web Meeting Fall 2022
Tuesday, October 25, 2022

The Panel met via Videoteleconference, at 10:00 a.m. EDT, Dave Nerenz and Christie Teigland, Co-Chairs, presiding.

Present:

David Nerenz, PhD, Co-Chair
Christie Teigland, PhD, Co-Chair
J. Matt Austin, PhD, Assistant Professor for Patient Safety and Quality, Johns Hopkins Medicine
John Bott, MBA, MSSW, Manager, Healthcare Ratings, Consumer Reports
Daniel Deutscher, PT, PhD, National Director of Research and Development, Maccabi Healthcare Services
Marybeth Farquhar, PhD, MSN, RN, Executive Vice President of Research, Quality and Scientific Affairs, American Urological Association
Jeffrey Geppert, EdM, JD, Senior Research Leader, Battelle Memorial Institute
Laurent Glance, MD, Professor and Vice-Chair for Research, University of Rochester School of Medicine and Dentistry
Joseph Kunisch, PhD, RN-BV, CPHQ, Vice President of Quality Programs, Harris Health
Paul Kurlansky, MD, Associate Professor of Surgery/Associate Director, Center for Innovation and Outcomes Research; Director of Research, Recruitment and CQI, Columbia University, College of Physicians and Surgeons/Columbia HeartSource
Zhenqiu Lin, PhD, Director of Data Management and Analytics, Yale-New Haven Hospital
Jack Needleman, PhD, Professor, University of California Los Angeles
Eugene Nuccio, PhD, Assistant Professor, University of Colorado, Anschutz Medical Campus
Sean O'Brien, PhD, Associate Professor of

Biostatistics and Bioinformatics, Duke University Medical Center

Jennifer Perloff, PhD, Scientist and Deputy Director, Institute of Healthcare Systems, Brandeis University

Patrick Romano, MD, MPH, Professor, University of California Davis

Sam Simon, PhD, Senior Researcher, Mathematica Policy Research

Alex Sox-Harris, PhD, MS, Associate Professor of Research, Department of Surgery, Stanford University

Ronald Walters, MD, MBA, MHA, MS, Associate Vice President of Medical Operations and Informatics, University of Texas MD Anderson Cancer Center

Susan White, PhD, RHIA, CHDA, Administrator - Analytics, The James Cancer Hospital at The Ohio State University Wexner Medical Center

NQF Staff:

Elizabeth Drye, MS, SM, Chief Scientific Officer

Tricia Elliott, DHA, MBA, CPHQ, FNAHQ, Senior Managing Director

Matthew Pickering, PharmD, Senior Director

Hannah Ingber, MPH, Manager

Gabrielle Kyle-Lion, MPH, Analyst

Also Present:

Rachel Brodie, Purchaser Business Group on Health

T.J. Christian, Abt Associates

David Gilbertson, Chronic Disease Research Group

Kathy Lester, Kidney Care Quality Alliance

Kristen Mcniff Landrum, KM Healthcare

Consulting

Feifei Ye, Rand Health

Contents

Proceedings

(10:01 a.m.)

Welcome, Roll Call, and Disclosures of Interest

Dr. Pickering: All right, we'll go ahead and get started. So good morning everyone. Once again, my name is Matt Pickering. I'm a Senior Director here at National Quality Forum.

And, welcome and thank you all for, especially our SMP members who are here convened to go through the, Fall 2022 Measure Evaluation for the scientific acceptability of the measure submitted for Fall 2022.

So I wanted to thank everyone for your time, especially those members of the public, and developers, being on the call today to answer any questions that the SMP members do have.

Also wanted to thank our SMP members, as well, for all of the work that you continue to do, and all the support you provide to NQF in our evaluation process.

This meeting today, but also leading up to this meeting, and your evaluation of those measures.

And, we look forward to discussing some of the measures that you've already evaluated today, and potentially re-voting on some of those areas that had some initial concerns.

Again, the developers of these measures are on the calls today, or on the call today, and they are there to answer any questions. And, we'll kind of go through that flow and process here in a little bit.

So we go to our next slide.

It's just a few housekeeping reminders. So we were just joking a little bit earlier, to say that we like, and do act like to keep our SMP members and others that join in our meetings, just up to speed with a lot of the technology in this virtual environment.

So we are now using the Zoom platform, so you may have seen that that's, may have been new joining into this meeting.

But it's a Zoom platform, but it has all the same features as some of the other platforms we've used before.

So you have the video feature, and of course, the microphone feature. You're able to call in, as well.

We encourage you to use your video, especially if you're chatting, or talking, just to be more engaging with this platform.

There is the chat box as well, which the staff and our co-chairs will be monitoring when we get to those discussions.

And then there's the raised hand feature, as well. And so when we get into those discussions if you'd like to raise your hand, we will definitely recognize you and you can participate in those discussions.

So we will do a roll call here, as well, at the beginning of this meeting just to establish attendance, but also a quorum.

We do recognize that some folks will be joining a little bit late, so we may do another check in on who may have joined late, before we get into the discussions of the measures.

If you have any technical issues, as always please don't hesitate to chat us in the chat.

Or you can email us as well, if you're having technical issues with the platform itself, at methodspanel@qualityforum.org.

Going to the next slide, please.

Just a few more here. We do have some meeting breaks built into our agenda today. We do have a lunch break, as well. That is currently scheduled around 1:15 for lunch.

We will keep to our reconvening at 2:00 p.m. just because folks, including our developers, are anticipating to join around the 2:00 p.m. time to discuss the Subgroup 2 measures. As well as members of the public may be looking to join around that time.

So even if we end a little bit early this morning, we will still reconvene at 2:00 to finish out the rest of those, those measures.

So just wanted to note that.

We also obviously have to recognize quorum or maintain quorum for voting, so if you have to step away for any reason, please message the team, or you can direct send a chat in the chat feature just to let us know you'll be stepping away, and for how long, so we can just make sure we keep an eye on quorum.

Because we need to keep quorum to vote, as you all are aware of that. So please let us know if that's, you're stepping away for any reason.

There is that chat feature I can mention, and the raised hand like I mentioned. The muting and unmuting is that little microphone button.

I think if you're calling in on the phone, there may be some instructions around hitting that *6 to take yourself off mute. So just keep in mind on that.

And if you're not speaking, we kindly ask for yourself to keep yourself on mute to prevent any background noise.

And, please, and introduce yourself when you're first speaking, just so we have a transcription service in attendance today that will be just tracking, and writing the transcripts for the meeting.

Next slide, please.

So now we're just going to go through welcome, introductions, and disclosures of interest.

Next slide.

I just want to recognize our two co-chairs, Dave and Christie, and allow them to give some welcoming remarks for today's proceedings.

So we'll first go to Christie, and then Dave.

Christie?

Co-Chair Teigland: Hi, good morning everyone. I, welcome to our fall evaluation review. I'm not sure where the summer went, but it went quickly.

So I hope all of you are getting back out in the world to attend some meetings, and conferences. I have been, and it's really such a pleasure to interact with people face-to-face again.

I didn't realize how much I missed that. So hopefully at some point in the future, we'll all be back together again.

But in the meantime, we're Zooming and it works. We've gotten pretty good at it.

So as usual, we have some really new and challenging issues presented to us in this round of measures under review for endorsement.

Just want to acknowledge again the, the NQF staff. They've done an amazing job synthesizing just vast amounts of information, and shepherding these new measures through the scientific review process.

I just can't overemphasize the amount of work that goes into this process when we kind of see a little bit from behind the scenes.

All the detailed materials they put together just to make all of this go smoothly for us, and seamlessly. So thank you to the staff.

But we're also really fortunate to have the Scientific Methods Panel being comprised of, we have experts from so many different fields of expertise, and disciplines that we know we need to evaluate all aspects of these complex submissions.

So you know, we always say that measure development is an art and a science, and fairness, you know, requires paying attention to all aspects. Not just those issues that are in each of our respective domains of expertise.

So we're constantly learning from one another. We're adding to the knowledge base of evaluation.

And so I really look forward to a lively discussion today, as we tackle some new issues.

Thank you.

Dr. Pickering: Thank you, Christie.

Dave?

Co-Chair Nerenz: Yes, thanks Christie, not much to add there. I just want to repeat the thanks to everyone who spent time and energy getting us to the point we are, as we started our work this morning.

NQF staff as Christie said, incredible work. The methods panel members, for the timespan and reviewing and discussing, and voting, and now the commitment of time today.

The developers for the work that they've done. And, the responses we received since the initial review.

All really important. The SMP role is very important in the NQF development, or endorsement process. That process is important in the overall scope of quality measurement.

So it's worth the effort on everyone's part, and I look forward to the work we're going to do right now.

Dr. Pickering: And, thank you Dave.

And again, just again thank you to the SMP members, as well as our developers, and other stakeholders for leading up to today's meeting and being on the call today.

We'll go to the next slide and just a, sort of a recognition as well of the SMP team. Here you can see on the slide Dr. Elizabeth Drye, the Chief Scientific Officer.

Trisha Elliott, our Senior Managing Director. Myself, Senior Director, as well Poonam Bal, who has supported this effort, as well, as a Senior Director.

Mike DiVecchia, our Director or Project Manager here, as well as Hannah Ingber, our other Content Manager and Gabby, as well, who is our Analyst.

So thank you to this great team.

And, going to the next slide.

So now I'll turn it over to my colleague, Trisha Elliott, who is on the call today to go over the introductions and roll call, as well as the disclosures of interest.

So Trisha, I will turn it over to you.

Ms. Elliott: Thanks so much, Matt. And I echo the thanks to all the committee members, and the NQF staff for getting us to this point, and to have this very valuable meeting today.

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

One is our annual disclosure of interest, and the other is disclosures specific to the measures we are reviewing in this cycle.

In those forms, we asked you a number of questions about your professional activities. Today we'll ask you to verbally disclose any information you provided on either of those forms, that you believe is relevant to this committee.

We are especially interested in grants, research, or consulting related to this committee's work.

Just a few reminders before we get started. You sit on this group as an individual. You do not represent the interests of your employer, or anyone who may have nominated you for this committee.

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

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

We'll now go around our virtual table. I'll start with the committee co-chairs. When I call your name, please state your name, what organization you are with, and if you have anything to disclose.

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

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

First up, I'll start with our committee co-chairs. David Nerenz?

Co-Chair Nerenz: Hi, Dave Nerenz, Henry Ford Health in Detroit. It's a large organization that is affected by quality measures in various ways, occasionally involved in development.

So in the past once in a while there has been a conflict, but on this particular cycle, no disclosures of any kind.

Ms. Elliott: Excellent, thank you.

Christie Teigland?

Co-Chair Teigland: Hi everyone, Christie Teigland. I am with the Novalon, Vice President of Research Science Advanced Analytics there.

We also do some quality measure development work with our health plans, and life science organizations.

But nothing to disclose in this round of reviews.

Ms. Elliott: Thank you.

Matt Austin?

Member Austin: Yes, good morning, Matt Austin. I'm an associate professor at the Johns Hopkins University School of Medicine, and I have nothing to disclose.

Ms. Elliott: Thank you.

John Bott?

Member Bott: Hi, John Bott. I'm an independent contractor currently helping in The Alliance in Wisconsin, and the Leapfrog Group.

And, I have nothing to disclose. Thanks.

Ms. Elliott: Thank you.

Daniel Deutscher?

Member Deutscher: Hello, I'm Daniel Deutscher, a research scientist at Net Health in the U.S., and also at the MaccabiTech Institute for Research and Innovation in Israel.

And, I have nothing to disclose for today.

Ms. Elliott: Thank you.

Marybeth Farquhar?

Member Farquhar: Good morning, I'm Marybeth Farquhar. I'm the Executive Vice President for

Research, Quality, and Scientific Affairs for the American Urological Association.

We do develop measures here but with regard to this slate of measures, we don't have any, I don't have any conflicts.

Thank you.

Ms. Elliott: Thank you.

Jeffrey Geppert?

Member Geppert: Jeff Geppert, Battelle Memorial Institute. Nothing to disclose for today.

Ms. Elliott: Thank you.

Laurent Glance?

Member Glance: Hi, good morning, I'm Larry Glance. I am a professor and Vice-Chair for Research at the University of Rochester, in the Department of Anesthesiology and Perioperative Medicine. I'm also at RAND Health.

I am on the American Society of Anesthesiologists Committee of Performance and Outcomes in Measures, the measurement.

We do develop measures, but none of these are before the panel today.

Thank you.

Ms. Elliott: Thank you.

Joseph Hyder?

(No audible response.)

Ms. Elliott: Sherrie Kaplan?

(No audible response.)

Ms. Elliott: Joe Kunisch?

Member Kunisch: Hi, good morning, Joe Kunisch with Harris Health System. I have no disclosures.

Thank you.

Ms. Elliott: Thank you.

Paul Kurlansky?

Member Kurlansky: Yes, hi, Paul Kurlansky, Associate Professor of Surgery at Columbia University.

I do sit on the Quality Measurement Task Force of the Society of Thoracic Surgeons, who does, who do submit measures to the NQF.

But I don't believe there are any measures that we're considering today. Otherwise, I have nothing to disclose.

Ms. Elliott: Thank you.

Zhenqiu Lin?

Member Lin: Good morning, my name is Zhenqiu Lin, I'm a Senior Director of Health Care and Analytics at Yale-CORE.

And, we develop measure for CMS and for this cycle, I think we submit two measure, but none of them were slate for today's discussion.

Ms. Elliott: Okay, thank you very much.

Jack Needleman?

Member Needleman: Hi, I'm a professor in the Department of Health Policy and Management at the UCLA School of Public Health, and I have nothing to disclose.

Ms. Elliott: Thank you.

Eugene Nuccio?

Member Nuccio: Good morning, I am a professor at the University of Colorado, Anschutz Medical Campus. I'm on inactive status for this round of measures, hence, nothing to disclose.

Ms. Elliott: Thank you.

Sean O'Brien?

Member O'Brien: Morning, Sean O'Brien, I'm from Duke University. I don't have any disclosures for today.

Ms. Elliott: Thank you.

Jennifer Perloff?

Member Perloff: Hi, Jenn Perloff, I'm a health services researcher at Brandeis, and I have nothing to disclose.

Ms. Elliott: Thank you.

Patrick Romano?

(Pause.)

Ms. Elliott: Oh, we can't hear you, Patrick. I think you're on mute.

Member Romano: Oh, yes.

Ms. Elliott: There we go, we can hear you.

Member Romano: Sorry, Patrick Romano. I'm a general internist in Health Services Research. We're based at UC Davis Health, in Sacramento.

We do develop measures on behalf of AHRQ and CMS. We have two measures in the current cycle, but neither are on the agenda for today's discussion.

Ms. Elliott: Thank you.

Sam Simon?

(No audible response.)

Ms. Elliott: Okay, Alex Sox-Harris?

Member Sox-Harris: Good morning, I'm a professor in the Stanford Department of Surgery, and a health services researcher, and research scientist at, in the VA system.

And, I have two federal grants focused on different aspects of quality measure science, but nothing related to the measures that we're reviewing today.

Ms. Elliott: Thank you very much.

Ron Walters?

Member Walters: Hi, I'm a medical oncologist, M.D. Anderson. I have no conflicts or disclosures.

I will say that anybody that's wondering, there's three medical oncology PROM measures, and I was not and am not, on the TEP that was involved in developing those measures.

That list, by the way, of who was involved, is available with a Google search.

Ms. Elliott: Great, thank you Ron.

Terri Warholak?

(No audible response.)

Ms. Elliott: Okay, Eric Weinhandl?

(No audible response.)

Ms. Elliott: Susan White?

(No audible response.)

Ms. Elliott: Okay, is there anyone that may have joined late, who didn't have a chance to speak up?

(No audible response.)

Ms. Elliott: Okay, we may have a few committee members that are joining a little bit later this morning. And, we'll have them disclose anything upon joining the meeting.

So thank you for that, I appreciate everyone participating in that. If you believe that you might have a conflict of interest at any time during the meeting today as topics are discussed, please speak up.

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

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

Send a message to your chairs, or to the NQF staff.

Does anyone have any questions, or anything you'd like to discuss based on the disclosures made today?

(No audible response.)

Ms. Elliott: Thank you for your cooperation with this aspect of the meeting.

As a reminder, NQF is a non-partisan organization. Out of mutual respect for each other, we kindly encourage that we make an effort to refrain from making comments, innuendoes, or humor relating to, for example, race, gender, politics, or topics that otherwise may be considered inappropriate during the meeting.

While we encourage discussions that are open, constructive, and collaborative, let's all be mindful of how our language and opinions may be perceived by others.

With that, I will turn things back to the team.

Thank you so much.

Dr. Pickering: Great, thanks, Trisha, and thank you all very much. As she said, we'll do a, just a check in in a little bit before we get into the measure proceedings, just to ensure that we know we have a couple people joining about 30 minutes late.

But we do have just some overview of the process and the measures, before we get into the actual voting and discussion of the measures today, so we'll, we'll go through that here just now.

So I'll turn this over to my colleague, Hannah, and Hannah will discuss the meeting overview today.

Hannah?

Meeting Overview

Ms. Ingber: Thanks, Matt.

So we've just gone through the welcome roll call and disclosures of interest. Next we'll go over the overview of the evaluation and voting process, just to get everyone on the same page before we start discussing the measures.

And, then we'll do the Fall 2022 measure evaluations. We'll break for lunch for 45 minutes and as Matt said, reconvene at 2:00 p.m.

Then we'll continue some more Fall 2022 measure evaluations, open up an opportunity for NQF member and public comment, and then go through next-steps for the rest of the Fall 2022 cycle.

After that, we will adjourn the meeting.

Next slide, please.

Just some meeting ground rules that we always go over at the beginning NQF meetings, and to add to what Trisha was just saying.

There's no rank in the room. We ask that everyone remain engaged and actively participate, and be prepared having reviewed the measures, and any additional materials beforehand.

We ask that members base their evaluation and recommendations on the NQF measure evaluation criteria and guidance.

That they keep their comments concise and focused, and be respectful and allow others to contribute, as well.

Of course, please share your experiences, and we look forward to learning from everyone during the call today.

Next slide.

For your meeting materials, one of the most important ones is the discussion guide. It's a synopsis document of the scientific acceptability content.

In other words, reliability, validity, and one composite construct for all complex measures in a measure cycle evaluated by the SMP.

Each measure includes pertinent information from the submission, SMP reviewer feedback, related developer responses, and identification of the measures that are pulled for SMP discussion during today's meeting.

The goal of the discussion guide is to summarize and highlight the important information for SMP discussion, and to reduce developer burden for multiple submission material or requests, and target critical scientific acceptability questions and concerns for discussion today.

Appendix B of the measure, of the Discussion Guide has additional information provided by the measure developers ahead of the meeting, for discussion today.

During our meeting, we also rely on some background materials that are linked in the slides, if you need access to them.

The first is the 2011 Testing Task Force Report. The 2021 NQF measure evaluation criteria and guidance, and the SMP-specific measure evaluation and guidance, which summarizes the pertinent information in the 2021 measure evaluation criteria and guidance document.

I'll now go through the overall overview of evaluation and voting process.

Overview of Evaluation Process and Voting Process

So the overall ratings are the same as always, but I'll go through your options. So, a high rating is available only if accountable entity level testing is submitted.

The measure may be eligible for high, but the sampling methods or results may warrant a moderate rating.

Moderate is the highest eligible rating if only a patient or encounter level testing, or face validity testing was conducted.

And again, a moderate might be what the measure is eligible for, but the testing may warrant a low rating.

Low is used primarily if testing results are not satisfactory, or an inappropriate methodology was applied.

Insufficient is used when the reviewer does not have sufficient information to assign a high, moderate, or low rating.

For example, unclear specs, unclear testing, or not conducting criteria required testing.

Next slide.

So, meeting quorum and achieving consensus is important for all NQF meetings. As always, our quorum is 66 percent of active SMP members in attendance.

Achieving consensus is calculated from the percent of quorum members in attendance during a vote, and is specific to the subgroup that evaluated the measure.

So, just to clarify, a pass or a recommendation to move onto the standing committee, is when greater than 60 percent of the yes votes, in other words, higher moderate ratings are received.

Consensus not reached is when 40 to 60 percent inclusive are yes votes. And, a no pass is when less than 40 percent of the votes are for yes.

So just some short reminders. All testing must align with specifications. If multiple levels of analysis are specified, each must be tested separately.

And NQF requirements permit passing some, or all levels of analysis for a measure.

Just some differences in testing requirements by measure type, as described in the measure evaluation guidance and criteria.

For reliability and validity, either patient or encounter level testing, or accountable entity-level testing is required for new measures. So, either or.

And that's true for outcome, intermediate clinical outcome, cost, resource use, structure, and process measures at initial submission.

For reliability and validity, testing is required at both the patient encounter, and accountable entity levels for instrument-based measures, at initial and maintenance submission.

Empirical analyses supporting the composite construction are required for composite measures, at initial and maintenance submission.

And if the patient encounter level validity testing is provided, we do not require additional patient encounter level reliability testing for any measures.

So we've broken this down by the requirements, and then which measures fall into that category.

So hopefully that's clear for everyone.

Next slide.

Just some additional reminders. Consideration for risk adjustment is required for all outcome resource use, intermediate outcome, and some process measures.

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

But concerns with discrimination calibration or the overall method of adjustment, are grounds for not passing a measure.

In the absence of risk adjustment or stratification for outcome resource use or intermediate outcome, and some process measures, a strong rationale or data for excluding, must be provided.

For all measures, incomplete or ambiguous specifications are ground for not passing a measure with an insufficient rating, as I mentioned before.

Next slide.

So what happens after the SMP review? The measures will then go on to the standing committees for their topic areas.

So standing committees will evaluate and make recommendations for endorsement for measures that pass the SMP review, and measures where the SMP did not reach consensus.

All measures reviewed by the SMP can be discussed by the standing committees.

Measures that don't pass the SMP may be pulled by a standing committee member, for further discussion.

For measures that don't pass the SMP and are pulled for discussion by a standing committee member, they may be eligible for a re-vote.

Eligibility for re-vote will be determined by NQF staff, and SMP co-chairs.

Measures that did not pass the SMP due to the following, will not be eligible for re-vote by the standing committee.

Inappropriately applied methodology or testing approach to demonstrate reliability or validity.

Incorrect calculations or formulas used for testing.

Description of testing approach results or data is insufficient for the SMP to apply the criteria.

Or appropriate levels of testing were not provided, or otherwise did not meet NQF's minimum evaluation requirements.

Next slide, please.

So I'll go through just the flow of today's discussion. Measures are discussed by the SMP in a pre-determined manner prior to the meeting, when we summarize the SMP's preliminary analyses.

The process for discussion during the meeting today for the measures that were pulled and on the agenda, is that staff will briefly introduce the measure.

The SMP member lead discussant will summarize their key concerns, and then other SMP subgroup members will be invited to comment.

The developers will be given two to three minutes for an initial response, and may respond to SMP questions during that time.

The discussion will then be open to the full SMP, and after the discussion ends, the SMP will move to vote on the relevant criterion, reliability, validity, or the composite construction.

The SMP voting process is conducted synchronously, virtually and confidentially via Poll Everywhere, as we've done in previous meetings.

Voting occurs following each criterion discussion.

SMP subgroup members only vote on the measures that they were assigned. And, recused SMP members cannot vote for measures where conflicts are identified.

Subgroup voting results taken during the meeting are the official SMP result, official SMP voting result.

And, for the measures that were not pulled for discussion but are in the discussion guide, those will pass in a consent calendar vote.

Next slide, please.

Are there any questions about the SMP evaluation process from SMP members? Or others?

Patrick I see your hand raised.

Member Romano: Yes, hi. My question is not really about today's discussion, but just for an update perhaps from you and Matt.

I know that we've had extensive discussion over the last couple of years about the criteria, and perhaps making the criteria a little bit more rigorous.

Particularly with respect to the reliability assessment, for example requiring both entity level, or requiring at least entity level reliability assessment for all measures.

So could we get an update about what the status of those proposals are, as they wend their way through the NQF process?

Dr. Pickering: Thanks, Patrick. So this is Matt on the call.

So we are still working through those updates for our criteria based on SMP input. Part of that as you know, is you know, convening with our CSAC to get their input, which we have done.

There's some additional work we're also looking to do, related to other aspects of our criteria. And, thinking about making updates going into next year.

On the table will be some of the SMP input that we received in the past year or more, related to reliability; related to requirements for empirical, empirical testing.

And even thinking about other aspects of our criteria, I think the scientific acceptability as you know, that we've been doing a lot of work right now with building out technical guidance for a social/functional risk adjustment.

So, that work will be concluding at the end of this year. So the next phase of that is trying to incorporate recommendations out of that work, into criteria.

So we're kind of folding that into the larger package of all of the other inputs that have come from SMP, but other, other aspects of our criteria may be updated.

So that work is ongoing. We'll definitely be ticking off more of that going into next year.

So currently at this point, we are still looking at the criteria that we've always been applying, which is that 2021 date that's on our website.

So, any other adjustments to that will happen going into next year.

So appreciate the question as I know I'm sure other SMP members are curious of where we are with the status on, on that.

But that's sort of the high level view of what we're going to be doing next year and some of the timeline, but I hope that answers your question, Patrick.

Member Romano: It does, thank you.

I mean obviously from the perspective, we've invested a lot of time and in that discussion, that process.

So I think many of us would like to see it move forward in 2023, into actual changes. But we understand there's a workflow that has to be accommodated.

Dr. Pickering: Yes, and thank you very much. And appreciate as always, the input we receive from the SMP, and all the great work that you've done with those threshold tables and changes to requirements as inputs for us, for us and our other panels to consider.

So we look forward to incorporating those and getting some more engagement with the SMP, especially during advisory meetings that we have with this group going into next year.

Any sort of updates to criteria, we also need to make sure that we provide an opportunity for public comment on things, as well as education.

Educational Webinars that we would host for developers and others, and standing committee members.

So, even though we make a quick change, it's not like a light switch. It would have to, you know, take its course as you mentioned, Patrick. We have this workflow that we will follow.

But that will be definitely going into next year, but appreciate the question, and thank you.

I do want to just take a moment because I know that the team has been, is messaging that a few other people may have joined.

So, I think Sam Simon, are you on the line? If you are, could you just introduce yourself, and would you mind just sharing if you have any disclosures, or any potential conflicts you'd like to disclose at this meeting?

Member Simon: Sure, happy to, and apologies for joining late.

Sam Simon, I'm a Senior Director at Mathematica. And, I don't have any disclosures.

Dr. Pickering: Great, thanks, Sam.

And, I believe also Susan White. If you're on, could you just introduce yourself? And mention if you have any conflicts you'd like to disclose?

Member White: Sure. Can you hear me okay?

Dr. Pickering: Yes.

Member White: Great. Susan White, I'm a Chief Analytics Officer at Ohio State Med Center, and I have a disclosure, and I'm having trouble finding the measure numbers.

I apologize for that, it's the PRO-PM measures that have been popular discussing. The 3720s.

Dr. Pickering: Right.

Member White: Okay.

Dr. Pickering: That's correct. Yes, Susan, so we have you down for the, those three measures which are 3718, 3720, and 3721.

Member White: I was pretty close, right?

Dr. Pickering: Yes, you were.

Member White: Good memory. Okay, thanks.

Dr. Pickering: However, I know that you didn't review those because you were in Subgroup No. 1, and those measures were in Subgroup No. 2.

But just for transparency, those were the disclosures, and thank you for that.

Member White: Thank you.

Dr. Pickering: Anyone else from SMP joined, that wasn't on the call during attendance?

(No audible response.)

Dr. Pickering: Okay, all right great.

Sorry, any other questions before we continue to our voting test?

(No audible response.)

Dr. Pickering: Okay, all right.

So with that, we'll continue moving forward today as we will just do a voting test, and then we'll go through our Fall 2022 measure overview.

And then we'll get into the discussions.

So, Gabby, I will turn it over to you.

Voting Test

Ms. Kyle-Lion: Thanks, Matt.

So I'm going to go ahead and pull up our, our test poll here. As a reminder, this poll is only for SMP members.

And SMP members, you should have received an email this morning with the Poll Everywhere voting link.

If you are having any trouble accessing the poll, please reach out to me, or Hannah, or Matt. Anybody on the NQF team will be able to go ahead and get that for you.

So I'll go ahead and activate our poll. So, the voting test poll is now open. The question is, do you like candy corn? Your options are yes or no.

And, we are looking for 19 votes here since Gene Nuccio is not an active member. We would have had 20 if he were. So 19 here.

And again, if you're having any trouble accessing the poll, please feel free to come off mute and let us know, raise your hand, or send a chat to any of the NQF team.

Co-Chair Nerenz: Gabby, what's the time of that email so we can find it?

Ms. Kyle-Lion: I believe I sent it around 9:15 this morning.

Co-Chair Nerenz: So I've got one with that timestamp, but it has join meeting. It doesn't mention a poll.

Ms. Kyle-Lion: All right, Dave, I can go ahead and forward you the email again. Just give me one second.

Co-Chair Nerenz: Okay, thanks.

Ms. Kyle-Lion: Yes, no problem.

Is anybody else having any issues accessing it, the poll?

Member Simon: The timestamp on my email is 9:12, if that helps anybody.

Ms. Kyle-Lion: Okay.

Member Simon: Close.

Ms. Kyle-Lion: Thank you, Sam.

Dr. Pickering: If you're on the Eastern side, it think it would have been 8:12. If that, yes.

Ms. Kyle-Lion: Dave, I just resent it to you.

Co-Chair Nerenz: Thank you.

(Pause.)

Co-Chair Nerenz: Got it.

Ms. Kyle-Lion: Perfect. I'm seeing 17. I'll just give it one more minute just to see if we get any other last minute votes here.

(Pause.)

Ms. Kyle-Lion: I still see us at 17, but that's okay because 16 is our quorum for voting. So, I'll go ahead and close our poll for now.

Member Kurlansky: What was the result?

Ms. Kyle-Lion: Twenty-eight percent of people said yes, and 72 percent of people said no.

So, looks like there's some consensus that candy corn is not good.

I will go ahead and pass this back to --

(Simultaneous speaking.)

Member Romano: I presume you'll notify us individually if our vote wasn't recorded?

Ms. Kyle-Lion: Yes, Patrick. If I didn't see your vote in there, I'll go ahead and message you all.

Member Romano: Thank you.

Dr. Pickering: All right.

Ms. Kyle-Lion: Sorry, Matt, give me one second to put back up the slides.

Dr. Pickering: No worries.

Fall 2022 Cycle Overview

Ms. Kyle-Lion: Okay. So actually this is me again, sorry.

All right, so I'm going to go ahead and take us through the Fall 2022 Cycle Overview. So go ahead and move to the next slide.

There were 13 complex measures assigned to the SMP. Of those 13, one was an outcome measure; two were composite measures; one was an intermediate clinical outcome measure; and eight were PRO-PMs; and one was a process measure.

Additionally, of those 13, eight were new measures.

So to evaluate those measures, we created two subgroups comprised of 11 or 12 SMP members, that were assigned six or seven measures for evaluation.

However, after the SMP preliminary review, two measures were withdrawn. NQF number 2881, and NQF number 2789, leaving a total of 11 measures remaining under SMP review.

Seven measures passed reliability, validity, and/or composite construction. Five measures total were slated for discussion today.

Four measures are slated for discussion due to receiving a CNR decision, or not passing on reliability, validity, and/or composite construction.

One measure was pulled for discussion, but is part of the seven that passed both reliability and validity.

This is just a, this slide shows a breakdown of the measures for discussion by subgroup, and the portfolio that they will be in after SMP review.

In Subgroup 1, there are two measures for discussion. NQF number 3725, which is assigned to the Renal portfolio.

And, NQF number 3654, which is assigned to the Geriatrics and Palliative Care portfolio.

In Subgroup 2 there are three grouped measures up for discussion, all assigned to the Patient Experience and Function portfolio.

And they are NQF number 3721, NQF number 3720, and NQF number 3718.

All right, with that I will go ahead and pass it back to Matt to start the discussion on the measures under review.

Dr. Pickering: Thanks, Gabby.

If you can go back to that previous slide just for a second, Gabby.

Thanks.

If you can see on Subgroup No. 2, 3718, you see the little superscript C and it indicates it was pulled by staff, or SMP.

So in this case, this measure it was pulled by staff just because of it is the same developer, the same type of analysis that was done for, for validity specifically.

And, so we wanted to pull it just for consistency.

So in the discussions today as we go through them, I think that the two measures that are being discussed for validity, are 3721 and 3720.

3718 did pass on validity, but the staff wanted to pull it just in case we needed to revisit that measure, and just making sure there's consistency in how we're applying the validity vote across all three.

Since it's the same measure, same developer, excuse me, same type of analysis for those.

So I just wanted to note that for the SMP's consideration.

In addition, if there's no need to revisit 3718 because of the issues with 3721 or 3720 are strikingly different, then there's no discussion needed with 3718.

But just wanted to make the SMP aware of that, specifically Subgroup No. 2 on why that measure is listed since it did pass on validity and reliability.

That will be for the afternoon.

So if there's no other questions, we can proceed, oh, I'm sorry, Gene you had put something in the chat.

So Gene, you had asked a little bit about the risk adjustment work. So just to answer that question quickly before we go into the next, or the measure discussion.

So that's correct. So there are seven what we call minimum standards from this risk adjustment TEP that are being proposed within the technical guidance.

That work again, is going to be concluding in December. So that technical guidance will be published and out, and been finalized in December.

The work that we have to do on the NQF side, is then translate those recommendations into our criteria.

So some of the things you're calling out Gene, in your comment here, could be considerations that we have to look at for reviewing these types of analyses for risk adjustment based on the requirements, or what those standards are within that technical guidance.

So as I mentioned, they're working that into the workflow for next year, and we'll be looking to engage the SMP as needed, on any of these types of changes.

So more to come going into next year, but Gene, thanks for the question. As the last TEP meeting we had was yesterday to go through some public comments on that technical guidance.

So we will be working after December, to start thinking about how to incorporate that into criteria.

So more work to come. Thanks for the question.

Member Nuccio: Thank you.

I was following up on Patrick's question about items moving forward, and I wanted to ensure that the work of the risk adjustment TEP workgroup, was integrated into our thinking regarding that TEP matter.

Dr. Pickering: Thank you very much.

Okay. No other questions. We will proceed. So we do have quorum for both of our subgroups, Subgroup 1 and Subgroup 2, so thank you all very much.

Measure Evaluation Subgroup 1

So we'll go into our first subgroup discussion, which is Subgroup 1. And for Subgroup 1 there are two measures on our agenda for discussion today. The first is 3725, which is Home Dialysis Retention. And the next after that would be 3654.

So before we proceed I just wanted to check in. Do we have a member from Kidney Care Quality Alliance on the call?

Ms. Lester: Hi, yes. This is Kathy Lester. I just want to make sure I'm joined by my colleagues Dr. Lisa McGonigal and the team from CDRG, Dr. Dave Gilbertson, Dr. Suying Li, and Dr. Jiannong Liu; I think I've seen everybody, so that we are here.

Dr. Pickering: Okay. Great. Thank you. And just as a reminder for the process flow for our developers, after the SMP Subgroup members are invited to comment or discuss any concerns that they have after a lead discussant concludes their comments, we then will have the developer give two to three minutes for a response to any additional comments that the SMP Subgroup has. So our chair will recognize a developer and then you'll be given about two to three minutes to give responses, just as a reminder.

Renal

#3725 Home Dialysis Retention (Kidney Care Quality Alliance)

Okay. So this measure, 3725, is the Home Dialysis Retention. This is the description for -- this measures the percent of all new home dialysis patients in the measurement year for whom greater than or equal to 90 consecutive days of home dialysis was achieved.

This is a new measure. It is an outcome measure at the facility level of analysis, and as I mentioned our Kidney Care Quality Alliance colleagues are on the call. They are the developer and steward for this measure.

The full description of reliability and validity for this measure can be found in the discussion guide on page 5, but today we are discussing reliability as the validity received a pass. So the reliability received a consensus not reached. And you can see the initial SMP Subgroup assessment votes there. And for the reliability, which I'll focus my summary on before turning it over to our chair Christie and then the lead discussant.

This measure was previously submitted to SMP under NQF No. 3697 as a clinical intermediate outcome measure. The developer resubmitted this as a clinical intermediate outcome measure under NQF 3725. And to account for previous feedback from the SMP the developer kept the level of analysis at the facility, but provided an explanation as to why the HRR level of analysis is not required. So you can see more also in the discussion guide.

But the reliability testing was conducted at the facility level using signal-to-noise analysis, the beta-binomial model specifically. The developer states that the HRR-level aggregation is not necessary for this measure because it only includes incident patients and does not need to account for facilities that do not offer home dialysis.

So the mean reliability at the facility level using one year of data was 0.604 with a median of 0.547. And the median facility had seven patients.

The developer noted that while the reliability statistics using one year of data meet NQF's criteria, they also calculated reliability by duplicating their data and treating it as a two-year rolling measure, given the small number of new home dialysis patients. So with those calculations the mean reliability increased to 0.846 and a median of 0.905 with the second year of data. And in their responses, developers' response they actually did some additional roll-ups as well doing three years of data, which I'm sure they can speak to if there are any questions related to those analyses.

They also performed additional analysis by randomly generating new yearly data for each facility and combined that with the 2021 data resulting in a similar increase in reliability. So 0.871 with a median of 0.931. And the developers argue that this analysis helps to alleviate concerns of auto-correlation.

So as far as what's on discussion for SMP is really to discuss and re-vote on reliability as it received a consensus not reached rating. So votes of lower or insufficient were due to the low volume units not obtained -- not obtaining adequate reliability. So votes of low or insufficient were due to the low volume units not obtained, not obtaining adequate reliability using one year of data as the measure is specified with one year of data as well as concerns surrounding the calculating used for the reliability score.

So with that summary I'll turn it over to Christie and we can kick off our discussion for the SMP.

Christie?

Co-Chair Teigland: Thank you, Matt. You gave a great summary of this measure and the steps that the Kidney Care Quality Alliance took to work through those reliability issues.

I'm just going to turn this over to Jack right now who is going to lead the discussion on some of the concerns that the committee had with the reliability testing and results.

So, Jack, please take it away.

(No audible response.)

Co-Chair Teigland: You're on mute, Jack.

Member Needleman: So Matt managed to take half of what I had sort of prepared (audio interference) comments here, so let me simply make a few points here: The basic structure of this and the measure that (audio interference) home dialysis (audio interference) are basically the same. The first measure of the patients eligible for home dialysis (audio interference) what percentage get it, take it.

Dr. Pickering: Jack?

Member Needleman: This one (audio interference) --

Dr. Pickering: Jack, sorry to interrupt.

Member Needleman: -- take it allowing for a 30-day training and testing period for whether people want to (audio interference).

Dr. Pickering: Jack, can you hear me?

Member Needleman: Yes.

Dr. Pickering: You're coming a little bit chopping. I don't know, maybe the video -- you might -- you could try turning the video off. That may help a little bit, but you're coming in a little choppy.

Member Needleman: Okay. Yes, I'm on -- okay, so is this better, Matt?

Dr. Pickering: I think so.

Member Needleman: (Audio interference) thinks it's better.

Okay. So basically what this is (audio interference) of those who start (audio interference) dialysis get through the (audio interference) period of 30 days and 90 days. And the denominator --

Dr. Pickering: Yes, sorry, Jack. You're still --

Co-Chair Teigland: You're cutting out, Jack.

Member Needleman: -- and that's where I think the reliability issues --

Dr. Pickering: -- kind of cutting out.

Member Needleman: -- emerge.

Dr. Pickering: Sorry, Jack.

Member Needleman: Okay.

Dr. Pickering: Yes, I don't know if you're logged on through the web link --

Member Needleman: Yes.

Dr. Pickering: -- but there may a phone -- try to call in through the phone. That may connect you a little bit better.

Member Needleman: Okay. Do you want to switch the order on the measures? Do you want somebody else to take this up?

Dr. Pickering: Yes, maybe if we can see if anyone else from the subgroup -- Christie, what do you think? Maybe (audio interference) subgroup wants to comment on any of the concerns.

Co-Chair Teigland: Anyone else on Subgroup 1 willing to comment on this measure, or shall we wait for Jack? I confess to not fully being able to articulate myself.

Dr. Pickering: Okay. So maybe, Jack, you want to try to call in? Let me just double -- let me check to see. Maybe we can go to our next measure just quickly as Jack is sort of connecting back in.

Do we have the developer for 3654 on the call, Hospice Care Index?

(No audible response.)

Dr. Pickering: Is Abt Associates on the call?

Dr. Christian: Oh hey there. Sorry, I had the had -- this is Dr. T.J. Christian of Abt Associates. We're here.

Dr. Pickering: Right, T.J. And sorry to our patient QA colleagues. We'll see if Jack can dial back in and maybe we can go back to that measure.

But for 3654, T.J., are you and your team okay to -- we switch to your measure and we'll come back to our patient QA measure?

Dr. Christian: I'm okay. Let me check if the steward is on. Looks like so. Yes, I think we're okay to go, Matt. Sorry.

Geriatrics and Palliative Care

#3654 Hospice Care Index (Abt Associates)

Dr. Pickering: Okay. Great. So we'll just do that. And sorry to our patient QA colleagues. We'll come back to that measure. We'll see if we can get Jack back up and running with a phone call. So thank you for your patience with that. Apologies.

And thank you, T.J. and Abt Associates as well. So you're on the call also, so similarly what we'll do is after the subgroup has some discussion around some of the concerns related to this measure we'll have our -- you and your associates at Abt Associates have two to three minutes to respond to any of the comments or concerns and then we'll proceed with a full group discussion after that.

So just going to 3654, our lead discussant is Sam Simon.

Sam, I know you're on the call. Are you okay to be the lead at this point?

Member Simon: Yes, good to go.

Dr. Pickering: Okay. Great. Thank you.

3654. This is the Hospice Care Index. So this Hospice Care Index monitors a broad set of leading claims-based indicators of hospice care processes. It reflects care throughout the hospice stay and by the care team within the domains of higher levels of care, visits by nursing staff, patterns of live discharge, and per-beneficiary spending. The index monitors ten indicators simultaneously, and compares individual provider scores to the thresholds which are set as benchmarks against the national distribution of performance scores. So hospices which are outliers are awarded a point for that indicator.

So this is a new measure. This is a composite measure at the facility level. And like I said, the developer is Abt Associates and our lead discussant is Sam Simon. You can find more information about this measure on page 9 of the discussion guide, but today we will be discussing reliability, validity, and also the composite.

So initially the reliability was a not pass as well as the validity, and the composite was a CNR. So as we proceed today we'll first discuss reliability and then there will be a re-vote on reliability before -- then we'll go to validity. I'll do the summary of validity. There will be a vote, et cetera. So we'll take it piece by piece.

so for reliability there was no component testing done, so none of the patient/encounter level testing done, but the developer for the accountable-entity level testing indicated that traditional approach of signal-to-noise ratio testing was not applicable to this measure and the developer instead conducted a stability analysis comparing index scores calculated for the same hospice using 2017 and 2019 data. So no statistical tests were actually conducted, but 46 percent had the same score in 2017 and 2019 and 15 percent had scores that differed by 2 points or more. So the developer does state that the design of the index with its focus on identifying hospices that are outliers in several areas ensures its reliability.

So several of the SMP members noted that while developer's assertion is correct that a signal-to-noise test was not possible, there could have been other tests performed such as the test/retest analysis, and therefore the testing provided was deemed insufficient from some of SMP members' concerns.

So for discussion on reliability is really to discuss the developer's responses to those SMP concerns and determine if the information submitted is enough to warrant a re-vote for reliability criteria.

So with that, Christie, I'll turn it to you and Sam and we can kick off the discussions for reliability.

Co-Chair Teigland: Yes, I'll just turn this over to Sam. That's a great summary of this measure. It's a fairly complex measure, but I think easy to understand how it was constructed.

Yes, and, Sam, I think for the -- particularly for the folks listening in maybe touch on also the difference between test/retest, which Matt just suggested might be another approach versus the stability -- you know, We got the same results two different time periods -- and why that's different. But please, Sam, take it away.

Member Simon: Yes. Sure. Happy to. And, Matt, I agree that was a good summary.

A couple of things I wanted to point out in addition to the points Matt raised. So this is really an interesting composite measure. It uses a -- it's a set of claims-based indicators, but what I found curious is that it's somewhat of a mixture of what seems like structural indicators like nursing minutes per home care day as well as costs, a process measure or two thrown in there, and a few outcomes including discharges followed by death in the hospital. So it's definitely an interesting and sort of a compelling composite measure, but it does mix a lot of these different types of measures.

And what's notable is that for some of these outcome measure -- the outcome measures that are included, although I will say that the developer considers these utilization metrics -- so we could talk about that, but I think the point I wanted to make here is that those particular indicators aren't risk-adjusted and the composite itself is not risk-adjusted. And so I think that's definitely a validity concern that I had.

So as Matt indicated, they used the stability analysis looking at index scores between '17 and '19 with a pretty healthy sample of hospices, about 3,500 hospices. And so they -- right, the developer looked at stability in those scores and found that about half of the hospices have the exact same score between 2017 and 2019. And a part of this may be due to the way the composite is constructed, because hospice is going to -- they -- essentially as long as they are not in the worst 10 percent of performance they get a point. So you have to be something of a -- it removes points for being an outlier.

And this nature is actually reflected in something else that I thought was interesting but also a little bit concerning for me from a validity perspective in that there's a bit of a ceiling effect when you look at the overall distribution of this measure. So 70 percent of hospices had a score of 90 -- sorry, of 9 or a 10. And there were very few that had -- I think like two percent that had a score of -- that had a score below five I want to say. Or sorry, below six. So it's a very constrained distribution, again sort of raising some validity concerns.

And so, but going back to sort of their specific validity testing, they looked at correlations of the HCI scores with CAHPS scores, two different CAHPS scores. First was a percentage of rating the hospice as a 9 or a 10 of care -- as caregivers' ratings and then another measure that looked at the percentage of caregivers that would definitely recommend the hospice.

So they correlated those two measures with the HCI and they found frankly pretty weak correlations of like 0.09 and 0.12, respectively, for each of those two measures. These were significant given the numbers, but again the magnitude was lower than what we'd like to see.

And then they also looked at hospices -- they looked at the adjusted odds ratio for hospices that had a lower score in HCI, which is a score of seven or below. And they looked at the lowest CAHPS Star ratings and they found that there was an adjusted odds ratio of two, but it was somewhat broad confidence interval. But it didn't go below one, so that is significant. So there are some signals of validity here, but also some concerns.

And then I know we wanted to talk about the composite. This is where we couldn't -- where the group didn't reach consensus. And I had several concerns around the composite. There wasn't necessarily a -- or there wasn't an indication of whether the composite was intended to be a reflective or formative construct.

Co-Chair Teigland: Sam, can I stop you for a second? Can we get back to the process here? We really need to talk about reliability first and vote on that.

Member Simon: Oh, I'm sorry.

Co-Chair Teigland: Yes, just -- NQF staff will not know what to do here unless we follow the process. So can we reiterate the reliability issues --

Member Simon: Yes.

Co-Chair Teigland: -- and then open it up to the SMP to comment on those reliability issues? That was a general broad summary of the reliability and validity issues, but let's get back to reliability and get that one out of the way, move to validity. Then we can --

Member Simon: Right.

Co-Chair Teigland: -- talk about the composite, which is a little more complicated. Thank you.

Member Simon: Sure. Sure. So the only thing I'll add around reliability -- basically so the developer did respond to this fact that there was an item level reliability done. There was a recommendation or an interest from the Methods Panel around using test/retest. The developer's response essentially was I think that there was an interest or a willingness to do this, but no new analyses were presented. The rationale -- one of the rationales provided was that none of these indicators that are used in the index are quality indicators. So that changes as the developer center due to what could be true drift. It's not that hospices were sort of tracking to these as measures to be used.

So I would say no, I didn't see any new information provided around reliability from the developer. Some justification, but no new data and no new analyses.

So those are the points I had around reliability. I'm happy to move to voting unless -- or others from our group who might have other comments as well.

Co-Chair Teigland: Yes, for others who voted this low can -- who's -- who else has some comments? Alex, you raised your hand.

Member Sox-Harris: Yes, thanks and thanks for the summary comments.

So on reliability it's a -- we need some empirical test reliability, not just the stability analysis that was provided. And to me reliability testing is about measurement error. So that's some kind of comparison of -- even if it's test/retest you have to think that the true scores did not change within that time. You can't push out test/retest too far.

So the underlying true scores have changed, so it needs to happen pretty rapidly. So even a test/retest done on two years that are -- or years that are two years apart, we would need to think or be convinced that true scores would not change during that time. So there could be some of split sample reliability testing or something, but we need an empirical test of reliability directly in order to pass the measure in my opinion. Thank you.

Co-Chair Teigland: Anyone else?

(No audible response.)

Co-Chair Teigland: Not seeing any hands. NQF Team?

(No audible response.)

Co-Chair Teigland: All right. Then let me -- if there's no other comments from the committee that voted on this, let's turn this over to you, T.J., for some -- you have feedback on those comments?

Dr. Christian: Nothing in particular. I just wanted to really kind of thank folks for their review and input. I'm sure that I'm kind of speaking on behalf of the steward as well, I think we're really just kind of here to learn and think about where to go next.

So just I mean from the review and then comments today I think we have some really good next steps we could do in terms of better establishing reliability. It sounds -- so in addition to kind of what was mentioned here, kind based from the comments, and plus we had kind of talks amongst ourselves there was kind of a thought that we could -- although we couldn't do sort of classic signal-to-noise testing for the overall -- the index the way it was formulated, we realized we could do it for the individual components. That's something that we probably will be pursuing in the future as well as the test/retest.

I think it was in one of the comments. It might have been one of the people that suggested test/retest. There was a sense that we could. Whereas you couldn't do sort of signal -- the way our measure is constructed we couldn't do signal-to-noise sort of within provider variance, but we could do -- I think the test/retest formulation was sort of a signal-to-noise over time. So that could be sort of a variation for provider over time sort of relative to all providers.

And I think if that's the current formulation that -- certainly that makes sense and it's compelling. To be honest, it's just something that we kind of like thought of previously just because we kind of stuck to the -- sort of the classical formulation of that. But in a sense there's other kind of empirical tests we could do and I think probably will be pursuing in the future.

Just wanted to thank members for their view and suggestion on that in terms of reliability.

Co-Chair Teigland: So you did have a large number of hospices. Was there a reason you didn't think about doing the split sample testing with this? It seems like there would have been enough.

Dr. Christian: Yes, I mean it's -- I think this one is kind of breaking the mold a little bit. I'm sure we could because we have a national sample of hospices, essentially all hospices in the country, and the way our measure worked was just to be against national benchmarks. So I think we just kind of used the whole nation as a benchmark sample. To do just kind of a like random half, that's certainly something also we could implement. I think that would be of interest to someone we could show as well.

Co-Chair Teigland: Yes, I mean the point that Sam made about the individual items being -- this is about the kitchen sink, right? You've got process, structure, outcome, cost measures, which really makes this composite fairly complex for us even to think about and wrap our heads around.

All right. Any other comment from the SMP on this measure -- I mean and subcommittee? And if not, do we -- should we re-vote on this, Matt? What is the procedure here, or do we need to?

Dr. Pickering: Yes, so that would be up to the SMP. So we'll see if there are any other SMP members, not even just the subgroup, have any follow-up comments or questions --

Co-Chair Teigland: Okay. Yes.

Dr. Pickering: -- related to this measure. And then after that we'll just ask if the SMP would like to revote on this or not. So for that we can just say would there -- does anyone disagree? Or I guess we'll say would anyone from the subgroup want to re-vote on the validity measure? It just takes one person to re-vote.

So first we'll go and see if any other SMP member even outside of the Subgroup 1 has any comments, questions for the developer.

(No audible response.)

Dr. Pickering: I don't see any hands raised and no --

Co-Chair Teigland: No, not seeing any hands raised.

Dr. Pickering: And so for this since there was this discussion and response from the developer -- and thank you, T.J., for your response and just very much appreciate also just trying to take in the SMP input for future considerations.

So with that discussion does any SMP subgroup member from Subgroup 1 wish to re-vote on validity -- excuse me, reliability? So you can just raise your hand. You can come off mute. You can also just direct message one of the team members if you'd like, if you want to remain anonymous. Do you want to re-vote on reliability?

(Pause.)

Dr. Pickering: Last call.

(Pause.)

Dr. Pickering: Sorry, Jack. Well, since you're back on, maybe -- you have something in a chat. How is your audio? Do you want to speak a little bit more about the chat you just put in?

(No audible response.)

Dr. Pickering: Jack, are you there?

(No audible response.)

Dr. Pickering: Sorry, Jack. Still can't hear you. I take it your comment wasn't about re-voting on reliability?

Co-Chair Teigland: Jack is asking about do we need to think about what additional tests we would want for a measure like this other than what we discussed already.

So, and, Ron, thank you for your comment, too.

He said I think there's opportunities to explore those survey-type situations where test/retest is not feasible. Hospice is one of those, so -- bereavement is another.

Yes, Jack, did you have something to say?

Member Needleman: Yes, I'm back.

Co-Chair Teigland: Yes, you sound good to me.

Dr. Pickering: Sound good, too.

Co-Chair Teigland: Good.

Member Needleman: Yes. So no, I think that we have given the -- expressed our concerns in the --

Dr. Pickering: Jack, sounds like you're walking away from the phone.

Co-Chair Teigland: Yes, you were great.

Member Needleman: Okay. It sounds like we agree that we had incomplete information to assess reliability here, but I'm not sure we've given the developers the information they need to figure out what reliability tests would be acceptable to us. And that may require more conversation after the meeting.

Co-Chair Teigland: Yes. Yes, other than the split sample idea we heard I don't -- I didn't hear any other recommendations for the developer.

Any other SMP members have any ideas about this, our statisticians?

Member O'Brien: I'm a Group 1 member. I'll just mention things that I would like to see. I thought that -- I forget who was talking who made the same point, just that if you look at two time points that are close in time and make the assumption that true performance didn't change much over time, then that is giving you an estimate of a signal-to-noise reliability under that strong assumption that there's no changes in true performance over time, but that when you estimate a correlation that can be interpreted as a signal-to-noise, proportion of signal-to-noise. And then I would just definitely like to see reliability testing for the individual item levels. If we go on to discuss validity, I'd have some accountability comments that actually are some related to reliability at the individual item level, that they're classifying the providers based on whether they're in the 10th percentile or not without really calculating a top presentable or any statistical testing.

So to the extent that those -- who ends up in the 10th percentile may be just the play of chance. If there's a lot of signal, then it's really going to be driven by signal. But if there is a lot of chance, there can be phenomena where the providers with relatively smaller denominators are the ones that end up in the 10th percentile more frequently. And I think you can have kind of counterintuitive associations that lead to these differences being explained by exactly unequal denominators rather than signal variation. So that can be I think tested empirically by looking at reliability at the individual item level.

Co-Chair Teigland: T.J., is that possible do you think for you to test reliability for each of those composite measures that comprise it?

Dr. Christian: Yes, I think that's something we might not have come across, but that's something we plan to do in the future.

Co-Chair Teigland: Yes. Great.

Dr. Christian: So I'm glad that it would be helpful to the Panel.

Co-Chair Teigland: Yes, I think that would provide us a lot more information.

Anyone else have any feedback here?

(Pause.)

Co-Chair Teigland: Not seeing any hands.

So any Subcommittee 1 members who think we need to re-vote this, would change their mind based on this discussion?

(Pause.)

Co-Chair Teigland: Doesn't look like we do. So let's move on.

Sam, maybe just reiterate. We'll move onto validity.

You already sort of described it quickly, but what were the main concerns with the validity testing?

Member Simon: Yes, sure. So yes, I can just do this rather quickly. I mean basically the correlations were definitely on the small side.

Co-Chair Teigland: Yes.

Member Simon: I think the concerns were on the distribution existed for me. And then ultimately the lack of risk adjustment is another sort of concern, red flag around validity. So I'll stop there and see if any of my colleagues have anything else they'd like to share or elaborate on.

Co-Chair Teigland: Sam, do you think the risk adjustment should be done at the individual measures that comprise this composite, or the overall composite, or both?

Member Simon: Yes, it's a fair question. I'm not sure I would sort of expect this -- I don't know, I think I'd rather have a statistician respond to that, but I just know that -- I mean it just seems like some of these things that are included in the composite have an outcome -- there's -- I can see an argument for considering this as utilization, but some of the ones that lean towards death definitely feel like outcomes and seem like they warrant some level of risk adjustment probably at the item level. But again I would leave that to a statistician to make a more educated point about that.

Co-Chair Teigland: Jack, you have your hand raised.

Member Needleman: I do. Thank you. I was one of the folks who voted incomplete -- inadequate information. And it was an issue with this measure, but it was also an issue with other measures we've seen. And we ought to provide a little bit more guidance to the developers.

And my concern is we've got 10 components, which is fine, but we -- all we get is the cut points and we don't see the underlying data and the distributions. I don't know whether these distributions are tight so that there's actually very little differentiation on factor 1 or 2 or 3, or whether they're wide so that cut is capturing a bit average difference between the yes and noes. And I can't evaluate the appropriateness of including these 10 items in a measure where I don't see the underlying distribution of the data.

Co-Chair Teigland: Yes. Yes, makes sense.

Anyone else on Subcommittee 1 have any feedback here or comments? And if not, we'll open up to the full SMP for any other comments.

Dr. Pickering: Patrick, you mentioned something in the chat. Did you want to comment on that or ask your question?

Member Romano: Sure. I mean this is a question, not necessarily for the developer, but just for discussion, that one could argue that given that hospice eligibility is limited to those who are first of all Medicare-enrolled and second are viewed to have a life expectancy of six months or less. There's a more constrained role for risk adjustment. Perhaps risk adjustment isn't necessary if risk profiles are actually very similar across hospice entities.

So I was wondering -- I mean obviously in this case I don't know whether the measure developer presented evidence on that question, but one could argue that the variation in risk is so small that risk adjustment is not necessary.

Co-Chair Teigland: Yes. Other folks have commented in the chat that -- the consensus in the chat at least is that the -- it's the individual measures that should be considered for risk adjustment. And given the wide mix or types of measures that may be the best approach. Seems like it would be tough to risk adjust this composite given the complexity of the different measures that are comprising it.

Dr. Pickering: Yes, and this is Matt of NQF. I'll just state that that aligns with our guidance. So if you go to our guidance criteria that are on our website, page 54 sort of lays out the different types of -- or the expectations for testing. Whether it be at the composite level or the component level risk adjustment applies to the outcome component measures. So unless they're NQF-endorsed. So NQF-endorsed measures could be used in composites. And it would be assumed that they would be evaluated for risk adjustment if they're already endorsed. So just drawing attention to that, to the commenters in the chat, that that aligns with our criteria.

Co-Chair Teigland: All right. Yes. Any other comments on this measure?

Looks like you've got your work cut out for you, T.J.

Anyone want to re-vote on the validity of this measure? Raise your hand.

Dr. Pickering: I'm sorry, Chris. Maybe before we go there, T.J., did you have any comments on the validity discussion that you'd like to share?

Co-Chair Teigland: Oh, sorry.

Dr. Christian: No, it's okay, Christine. Thank you.

Just I guess really one kind of global thing just to really thank everybody. I know Jack in particular and others really kind of made a point of working to educate us. I know Christie said we do have our work cut out for us and we kind of realized that and had no other comments, and why we're here, but see ourselves I think committed to this measure and really, really interested in going back and improving it. And I think based on what we're learning today it's really helpful to understand the next steps we should take.

In regards to validity in particular, I think there are a couple of different areas: I'll just -- I know time is limited, so I'll mention them briefly. Sort of the overall correlations with the CAHPS scores. I think we agree with Same. It's weak. There's something there. So hopefully simultaneously with this where the Hospice Quality Reporting Program is a younger initiative for CMS. There's not as many outcome measures at present, but others are in development. So hopefully that could be another source of measures to test against and data to obtain.

We thought we were going to perhaps convene a group of -- like another TEP or other external experts to get their input on some of these measures, to get some face validity in addition to any empirical testing that we could do. Actually in regards with that group the point was raised about the thresholds and where -- we're trying to separate between the outliers verse the other often 90 percent of hospices. And I guess Jack and I think others raised the point that there wasn't a good indication of what those differences were and how meaningful they were. I think that's definitely a valid point. I think with hospice there's not really strong clinical guidelines for some of these things. There's not really a hard and fast rule. So our part was more to try and do more of a relational comparison, but it is valid. There could not be a really strong or meaningful difference.

So just getting some -- we can certainly show the data, but also getting a little bit of experts to confirm whether or not there's a strong difference we think would be helpful. And hopefully that would satisfy some of that.

I think lastly the risk adjustments came up. I think that's certainly something we could consider. Depending we do some of this testing, if we change sort of the components, we could see if there are still arguably outcomes or not. But I think we can certainly consider that as well.

I will say some of our other thoughts were that -- because we're trying to find the outliers. We weren't sure if our risk adjustment was going to, you know, actually get a hospice into the bottom 10 percent, but certainly on the line there could be some differences. So if there is some concern, we could ask a TEP or another expert panel as well if there's a possibility these should be adjusted. That's certainly something we could consider as well. So appreciate all those helpful thoughts.

Co-Chair Teigland: All right. Yes, that's great feedback and hopefully that -- some guidance from the SMP here as to where to go.

Okay. And I don't see any calls for a re-vote on this one, Matt, so good to move on.

(No audible response.)

Co-Chair Teigland: You're on mute, Matt.

Dr. Pickering: I was just agreeing with you, Christie. I don't see any hands raised for re-voting and --

Co-Chair Teigland: Yes.

Dr. Pickering: -- or wanting to re-vote in the chat. Looks like Sean is dropping in some additional -- so, T.J., Sean O'Brien is dropping in additional thoughts on maybe some testing, but still no call for re-vote on validity. So the vote of the not pass would stand and we can move to any composite discussion as needed.

Co-Chair Teigland: All right. So, Sam, let's move onto the composite measure, which interestingly was consensus not reached probably because of a lot of confusion about what these results sort of meant and if they were even valid reliable at the composite level. What are your thoughts on this one?

Member Simon: Yes. So I guess I go back to -- well, let's talk about what the developer did for their composite.

So there wasn't a lot of empirical analysis of the composite. The composite was reviewed by a TEP, by the project TEP. And so when you look at NQF guidance around what is required or what is -- what the NQF guidance is, I believe the guidance is pretty clear that there should be a systematic assessment of content or face validity of the composite. And so if that was done, if there was something systematic about the review by the TEP, it wasn't clear from the materials, either in the original documentation or the response from the developer.

The other thing I was thinking about the composite is that we don't know if this was as a -- intended to be a reflective or formative composite. That would give us -- or even a logic model for this composite. I think that would have been helpful for us as we evaluated the composite approach.

And then I think many of us -- I mean I was curious why there were no sort of correlational analysis provided. I think that could have been something that the developer could have presented if this was designed to be a reflective composite.

So I think that there was some statistical testing; I don't mean to imply there wasn't or was some, where the developer looked at the standard deviation for each of the indicators as well as for the entire composite sort of systematically removing each item. And so the standard deviation was the broadest for when all of the indicators were included. I didn't find that particularly compelling.

So I think those are those are the main points I wanted to make. Again so I'll stop here and let others make their points as well.

Co-Chair Teigland: Any other Subcommittee folks who have any comments on this composite measure or vote?

(Pause.)

Co-Chair Teigland: Seeing no other comments from anyone here. So I think that's a wrap.

Matt, I think the -- since this one is no pass/no pass, the consensus not reached, will this move onto the Standing Committee? What's next?

Dr. Pickering: Yes. So I'll see also -- T.J., do you have any comments you'd like

to --

Co-Chair Teigland: I keep forgetting to ask T.J. I'm sorry.

Dr. Christian: It's okay.

Dr. Pickering: -- provide on the composite?

Dr. Christian: I'm pretty forgettable. No, I guess for the last time just to echo thank you. We're really privileged to be at the front of many smart people. And we just -- we were kind of talking, I and my colleagues were kind of talking amongst ourselves. Really a lot of great ideas to try. Hopefully we'll see you again soon with a new and improved measure. And just -- again just really thankful for all the input and comments towards this.

Dr. Pickering: Okay. Thanks. Thanks, T.J.

So at this point again since this is another criterion under the validity assessment for this measure we just wanted to check in. So right now it's a CNR. The SMP can determine they don't wish to re-vote on this composite, again similarly like we haven't been re-voting on reliability and validity. If so the CNR will be -- will hold on this for the composite.

And then the Standing Committees will be able to review. And they can always pull a measure for discussion. And we'll have to look at the eligibility if it's eligible for a re-vote on reliability and validity, but at least the Standing Committee will know there was a CNR on the composite assessment coming from the SMP.

That's if there's no one that would like to re-vote.

But I do see a few questions that have popped up, or at least hands raised. So Z.Q. and then Jack.

Member Lin: Just a question. When I look at the slide 28 -- so for reliability is 3-8, no pass. For validity it 3-8, no pass. For composite is 5-5, consensus not reached. I mean I can easily see it's not 5-5, right? Six-four. So how do we handle that? Like if one measure doesn't pass reliability or validity, do we move onto the composite?

Dr. Pickering: That's a great question. So we do want to ensure that the SMP has discussed or considered the CNR, just as we've been doing today, just reviewing the assessment of CNR, having any discussion, as well as having the developer provide any responses. If there's no need to re-vote on the composite because of the concerns with validity and reliability for example, or there's just no need based on the developer's response, the SMP may decide not to re-vote and the CNR will just stand. It will go to the Standing Committee if they'd like to pull the measure for discussion, but that is where it would reside. So if the SMP does not wish to re-vote on the composite and just leave the CNR based on the discussions today, that is an option.

Co-Chair Teigland: Yes, we discussed that, Z.Q., at length in evaluating this measure. Should it have even gotten to the composite given it didn't pass reliability or validity? That's not part of the current rules as were not told by the NQF staff. I'm not sure how we -- the consensus not reached got there given the no pass/no pass, but that's where we landed. So maybe -- I mean if there is any Subcommittee member who would like to re-vote the CNR their vote on the composite, we could do that, I guess.

Right, Matt?

Dr. Pickering: Yes, we can do that.

Co-Chair Teigland: Yes.

Dr. Pickering: So if one member likes to re-vote, please just let us know. You could even direct message one of our team members, if you'd like to remain anonymous, and we can re-vote on the composite. But if no need, we'll keep it a CNR.

Co-Chair Teigland: Yes. Jack does have his hand up. I'm not sure if it's to re-vote CNR.

Member Needleman: No. No, it's not to re-vote.

Co-Chair Teigland: No? Just a comment? Okay.

Member Needleman: Just a more general comment. I made the comment that I found it very hard to assess the validity without seeing the underlying data about what -- about the measurement for each of the individual components.

This is an extraordinarily talented group with enormous experience and expertise in quality measurement. So I don't want to call us generalists, but we don't necessarily have specific expertise in understanding hospice care. And if the Standing Committee pulls this one up for whatever -- for any reason, they do have more expertise than us to evaluate the individual components of the composite. But they can't do that without the underlying data beyond what where the cut point was in the distribution.

So, T.J., if this does get pulled up to the Standing Committee as well as for the next iteration, I would just encourage you to augment what you've already got here with information about the underlying data in each of those composites to make the case that each element of the composite deserves to be in a measure of hospice care and that there's enough variation there that it's important to capture that in a did -- are you good enough on this point?

Dr. Christian: That's helpful. Thank you, Jack.

Co-Chair Teigland: All right. I think we captured everyone's thoughts on this, Matt. We'll let the CNR rating for the composite stand and move to the Standing Committee if they decide to review it.

Dr. Pickering: Sounds good.

All right. Well, T.J. and our colleagues at Abt Associates, thank you very much for attending the call today. I obviously want to welcome those and everything else, but thank you for also kind of being flexible in going first as well.

We'll conclude that discussion. So all of those votes will be sustained here, just there's no re-vote as the SMP has not re-voted on all of these. So these votes will stay. And then we'll follow up with next steps with this measure after the call, which includes sharing the votes with our Standing Committee.

Okay. So I want to switch back to the other measure that we originally started with, which is 3725. So again that's the Kidney Care Quality Alliance.

So again, I'm just going to check with our developers there. Members from our Kidney Care Quality Alliance, are you good to go for this measure?

Ms. Lester: Yes, Matt, we are. I think we've got everybody still on the line.

Dr. Pickering: Thank you so much for your patience and flexibility there. Thank you so much.

Ms. Lester: Not a problem.

Dr. Pickering: So we already did the summary. And again we're talking just on reliability since validity did pass.

And, Jack, since you're coming in quite beautifully, I will go ahead and turn it over to you for your summary.

Member Needleman: I want to thank both sets of developers for their flexibility. I want to apologize for my technical problems.

I think on this measure the reliability summary that Matt provided accurately captured the data. Just to remind people about what this measure is, we start out with patients in dialysis centers and their -- there's an effort to get people on home dialysis, so you want to see what percentage of the folks who are eligible wind up on home dialysis. That's the measure that passed.

And then you want to keep people on dialysis. So this is the retention measure, which basically says for those who started home dialysis and got through an initial 28-day training and testing period and decided to continue with the home dialysis after that initial orientation and testing and trying it, did they have 90 days of continuous home dialysis? So it's a continuation measure.

The denominator is much smaller than in the first measure because you only get people in this measure who are -- who have started home dialysis. So we start with a small denominator and that's where a lot of the reliability issues emerge.

The initial set of reliability estimates are as Matt described with one year of data, a mean of 0.6, which under our new standard still be subject to review just crosses the threshold and a median of about 0.547, a little bit lower. And for a lot of -- which means for half of the distribution on reliability as well. And I think that's where the Committee reviewing this initially reacted. It just doesn't -- with the small Ns it just doesn't feel reliable enough.

The developer came back and said well, we could go for a two-year measure. And they simulated what would happen with two years of measurement in a couple of ways. And under their estimates those numbers went up. The mean reliability was 0.84; the median reliability was 0.9. And those look like really solid reliability estimates.

What we don't know because it hasn't -- I haven't seen anything formal is whether the developer is now sort of amending their thing to say we would only recommend this be used with two years of rolling data, or three years of rolling data.

So I need to hear that, Kathy, when you respond because I think that's the bare minimum for hitting the reliability threshold.

One of the reviewers who looked at this however raised other reliability concerns, and I think it's important to air those. And I hope whoever made the comment will share it.

Basically what they said is the whole signal-to-noise measure is dependent upon having a decent estimate of the probability, the estimate of P at the provider level, at the entity level. And when you have small Ns the estimate of P is also unreliable. And what they said was they thought that with the small denominators the P had is not reliably estimated, so the true variance can be over or underestimated with underestimation more likely, and unestimation means the reliability estimate is too high.

And they did some simulations in their comments where they basically pulled all of the Ns down to basically the -- where the range was and they said even with the two years of data their estimated reliability at the median was 0.39. And if that's an accurate -- that was not shared -- well, it probably was shared with the developer but was not part of the presentation but it is one of the concerns that's been raised in comments on the measure. And 0.39 under our getting away from biometrics as arbitrary numbers is a low level of reliability, not one that we feel comfortable endorsing.

So I think we need to talk more about the pooling. We need to talk about whether this is going to be a two-year measure to your rolling measure. And frankly, I'd like to hear my colleague talk a little bit more about their discussion of the unreliability of the

Ps and therefore the potential overestimation of reliability even in the pool of data.

Dr. Pickering: Sorry Christie, I think you're on mute.

Co-Chair Teigland: Thank you, Jack. Yes, I saw I was on mute.

And I'm not seeing any hands from our Subcommittee to comment further on that.

Yes, I'd like to hear a little bit more about how they simulated the 0.39 rating. I'm not sure our Kidney Care folks fully understand that from your description, Jack. Maybe that --

Sean? Thank you.

Member O'Brien: Yes, you may have been referring to my comments and I think it gets a little bit technical. My impression from the developer's responses is that they read some of the comments I provided and did calculations that kind of addressed my concerns.

The bottom line, I think, they've done something that that kind of made sense at least to me and it could be something to discuss in a later session. But there is -- the formula and the RAND tutorial that they reference a lot, it's defining reliability either in the ratio, the proportion of signal variation, but the total variation, but it has an interpretation of a squared correlation, Pearson correlation between the estimate and the truth.

And if you literally try to apply that definition you have different denominators across the sample sizes, you can't really calculate that correlation. And so there's kind of some formulas in there that try to approximate reliability by plugging in estimates for each individual provider based on an estimate of each individual provider's different provider variation. That different provider variation depends on probabilities that we can actually observe and we can estimate them, but when you have various small denominators, you can have probabilities of estimated of 100 percent or zero percent probability. And that phenomenon -- and that if you get an estimate of 100 percent, then it turns out your estimated within provider variability is zero which isn't likely something we believe -- and as an overall phenomenon if you kind of study it from a probability perspective, there turns out to be the systemic bias where you can overestimate reliability.

So I've proposed another formula that if you say let's define reliability as the squared correlation between the true measurement and the estimate, if I really have the same denominator, what would the formula for that quantity be? And it has a little expression and then you can evaluate that and plug it in with any sample size. And that's asked the question will the reliability be if everybody had ten cases or what would the reliability be if I had 20 cases, et cetera. And that was the suggestion and it looks like they implemented that.

And I haven't actually -- I mean it's not something that I've seen done. It's something that I've done for measure submissions that I've participated in myself. So I don't have a great justification or reference for it.

Co-Chair Teigland: Great. Any other Subgroup 1 members who -- Alex.

Member Sox-Harris: Thank you. I have a much less technical concern. I appreciate Sean educating us and I'd love to see some of the details of that.

The measure, as stated, is a one-year measure. And so I get stuck on that. They did some supplementary analysis showing that if it was a two-year measure, reliability might go up. But it's currently specified as a one-year measure. So that's my main decision point.

I would also say if it was a two-year measure, we would need two years of data to work with for the analysis because just simulating out of one-year data I think is going to overestimate reliability just because -- just assuming that two years is just a -- things don't change much.

So methodologically, I would have problem with testing a two-year measure on re-use of one year of data. So thank you. Those are my concerns.

Co-Chair Teigland: No, that's very clear, Alex, and that makes perfect sense and the developers might want to consider when they resubmit. Because if you don't specify that it's a two-year measure, in practice it won't be used that way which means it may not be reliable in practice and that's a concern.

Any other SMP members outside of the committee that would like to comment on this?

If not, we'll turn to Kathy. Would you like to provide some feedback on what you heard so far?

Ms. Lester: Yes. Thank you. And we very much appreciate the opportunity to submit the measures and the thoughtful review of the measures.

Focusing in on sort of what has been raised today, I do appreciate that the SMP really understands the importance of this retention measure. I thought Jack's summary was very good.

From our perspective, this is a must-have measure from the patient point of view. They don't want to be thrown on to home dialysis only to come off of it. So the retention measure is a guardrail as we've talked about before.

And just to sort of reiterate, this is a measure when you look at the specifications, it uses the term measurement year. CMS assigns a measurement year through its implementation policy. So it is not, in fact, limited to a single year and this is a measure that we don't intend to implement, but that is intended to be submitted to CMS to be part of its ESRD Treatment Choices program.

So the two recommendations that we had in there for implementation, you know, really go to how CMS would do that. On the small number side, I think for those of you who have looked at CMS ESRD measures before, you'll note that they had a systematic exclusion for in each facility that has fewer than 11 patients. So that's automatically going to be applied whether we outline that or not.

And similarly, when you look at how they define the measurement year, they can adjust that. So I don't think that we have to be locked into a single year and as you saw in the response looking at a three year rolling time frame is what we would propose to CMS as the best way to define that measurement year.

I'm going to turn to my colleagues at CDRG here to address Alex's concern, but just as a starter we do have a lot of history of dealing with small numbers in the ESRD. While it is a growing population, a lot of these folks, you know, it's a very small percentage of Medicare overall and so I think the idea of three year rolling time frames we've seen it in other areas before.

I would also say that when you look at dialysis data today and Dave is certainly the expert in the country running the US RDS program of dialysis data, we don't see a lot of flexibility or variability currently in that data. So one of the reasons we felt we could have a simulation of those data is that there has been relative consistency. We think over time that that will change and so that's why this measure is so important.

But Dave and maybe Sue, can I turn it to you to address Alex's concern about testing with simulated data for years two and three?

Dr. Gilbertson: Yes. I can take a shot at it. So the first thing we did was to just simply duplicate the data we had, understanding that that would most certainly sort of overestimate reliability. So then we thought we would simulate it.

Then to Alex's point that may be an overestimate because we can't count on things being exactly the same the next year. However, as Kathy mentioned, if we recommend a three year rolling average that in cases the reliability estimates significantly and so even if there may be year-to-year change somewhat, I would most certainly think that the reliability estimates would still be quite high, maybe not the 95, 99 percent, but certainly quite high.

Ms. Lester: So having answered the questions, happy to go into a deeper dive of anything else. I know there were some issues around the individual facility level versus the HRR, but it sounds like those have been addressed, so I will leave it there.

Co-Chair Teigland: I'm fine with that.

Ms. Lester: Thank you for that feedback. You're right, CMS definitely has some rules that they -- and terminology that they applied when they implement these measures and they very well could make this a two or three year rolling measure based on that with lots of measures and they always exclude, you know, less than 11 from denominator facilities or entities. That's a standard practice for all measures, right?

Any other feedback or comments from the SMP and if not, anyone on -- so this measure did pass on validity. It did not pass for reliability. Is there anyone who thinks we need to re-vote after hearing this discussion and the feedback from the developers?

Dr. Pickering: Just to confirm, there was a CNR on reliability and so it did not pass for this.

Ms. Lester: Oh, I'm sorry. It was CNR, yes. Some were close.

Dr. Pickering: Jack, I think he has his hand raised.

Member Needleman: Yes, and it's over the issue of re-voting. So this is a new measure. Kathy has said from the perspective of the kidney disease community, it is a critical measure, guardrail measure. We've got some encouragement to CMS to implement it with a long enough time frame beyond the standard -- what we all can see is the one year time frame to make sure the volumes are up and to assure reliability.

I have been frequently critical of CMS for picking low volumes to try to be more inclusive, but I want to believe that the message we've made about needing adequate numbers for reliability that Kathy echoed, will be effectively communicated to CMS.

And given all those things and the fact that it's a new measure and we will be seeing it for re-endorsement in a reasonable time period, I would be open -- I would like to consider a re-vote on the reliability at this point for purposes of moving it up to the Standing Committee.

Co-Chair Teigland: I would second that. I think a re-vote, given the discussion and the feedback from the developers would be a good idea, so let's -- can we move forward with that?

Dr. Pickering: Sounds good. Thank you. So we'll go ahead and move forward with the voting for this measure. So again, this is for 3725, the home dialysis retention. This is only for Subgroup 1 participants, so you're voting on just reliability as it was a CNR.

So I'll turn it over to Data and you can pull up the voting for 3725.

Co-Chair Teigland: Like Matt said, this is just for Subgroup 1 SMP members. Voting is now open for reliability on NQFM 3725. Reactions are A for high, B for moderate, C for low, or D for insufficient.

And I believe that we have eight votes for this, eight people present from Subgroup 1, so we are looking for eight votes here.

I'm seeing seven votes. All right, we just hit eight, so I'll go ahead and close the poll.

Give me one moment.

Ms. Kyle-Lion: The suspense is killing them.

Co-Chair Teigland: I know.

Ms. Kyle-Lion: I can see your face, Kathy.

Ms. Lester: I'm sorry.

Ms. Kyle-Lion: These are our babies.

Ms. Lester: I know.

Ms. Kyle-Lion: You've been there.

Co-Chair Teigland: Okay, sorry, everyone for that momentary pause.

Okay, so there was zero votes for high, five votes for moderate, two votes for low, and one vote for insufficient. Therefore, the measure passes on reliability.

I'll go ahead and pass it to you, Matt and Christie.

Ms. Kyle-Lion: Well, I see some happy faces from Kidney Foundation and this is a valuable measure. There's no question I think with the online SMP members that this is a much-needed measure. But our job is to make sure we're accurately measuring what we say we're measuring and I think as long as we have -- we'll have to push to CMS to -- but they understand those issues.

So great, I think we can move on, Matt. Thank you, all.

Ms. Lester: Thank you so much.

Dr. Pickering: Thank you very much. So we are way ahead of schedule, right? So we are now at 12 o'clock and we originally were thinking that we would go to 1:15, given if there was any further discussion on some of the areas of the hospice care index measure. So we're way ahead of schedule.

And I can mention we really do want to try to keep to the 2 o'clock hour just because we want to make sure the developers for those measures are going to be in attendance.

I'll just check in real quick. Is anyone from the purchaser or excuse me, the -- let me go down the list, Purchaser Business Group on Health, are you on the call?

Ms. Brodie: Yes, hi, can you hear me? This is Rachel Brodie. I'm the project director.

Dr. Pickering: Great. Hi, Rachel. So we are way ahead of schedule and we did break for lunch here, but rather than waiting until two, is it possible to come back a little earlier?

Ms. Brodie: Absolutely. Feifei Ye from RAND is also on the line and also Kris McNiff who is our -- Ms. McNiff Landrum who is our methodology specialist. So if you just tell us the time, we will all three be back on the line.

Dr. Pickering: Okay. Okay. So does the SMP have any concerns if we just break for an hour, we'll come back at 1 o'clock as opposed to 2 o'clock and then we'll pick back up with the last remaining measures for Subgroup 2.

Any concerns with that from SMP folks? Let me see if there any hands raised. Okay, that's great. So we'll definitely do that. We'll take 58 minutes now. So we'll come back at 1 p.m. Eastern and we'll pick back up with Subgroup 2 and close out the remaining discussions for those measures.

So 1 p.m. Eastern. Thank you all very much and we will take a break.

(Whereupon, the above-entitled matter went off the record at 12:02 p.m. and resumed at 1:00 p.m.)

Measure Evaluation Subgroup 2

Patient Experience and Function

Dr. Pickering: Okay, so we have 1 o'clock on the Eastern side which we are reconvening, so the recording has started again. So now we're going to go into the Subgroup 2 discussions, so these are going to be three measures, actually two measures, but the third one like I mentioned earlier, 3718, will be pulled just for any further decision making on consistency with evaluation. So if the SMP determines that 3720 and 3721 are strikingly different with how those validity assessments were, we won't need to go into 3718 unless the SMP wishes to do so.

We do have the developer on the line, the Purchaser Business Group on Health, for those three measures.

But before I get started I just wanted to revisit Dr. Romano's comment in the chat, so thank you, Dr. Romano, about confirming that all subgroup members who were on the line are voting. Originally, we did have nine for Subgroup 1 on the call, but one of our SMP members had to leave before the vote, so that did drop us down to eight.

So our apologies for not clarifying that on the call while the vote was going on and we did have one of our SMP members drop off the call, so that dropped our numbers down to eight, but that still was a quorum for that subgroup in which obviously you saw that the measure passed on reliability.

So thank you, Dr. Romano, for keeping us on our toes and apologies for not mentioning that earlier.

Okay, so we'll go into Subgroup 2 and it will be a very similar process as we've done previously with this group. So that being said, I'll summarize the issues related to the measure. We'll first start out with reliability during the reliability issue and then we will turn it over to our co-chair and our lead discussant to present any additional concerns and then open it up to the other subgroup members of that measure in which you also provided any comments or questions for the developer.

The developer is on the line, so after those comments or questions from the subgroup are shared, we will then have the developer provide a two to three minute response to those and we'll go back to the SMP for additional comments in which the entire SMP can ask questions and the developer is still on the line to answer those questions. And then we'll go to the next criterion if that is the case which is validity and so on and follow that same process.

So before we get started, any questions before we go into our first measure for Subgroup B?

Co-Chair Nerenz: Matt, if I could just take a minute and do a little framing with the group I think it might streamline our discussion.

Dr. Pickering: Certainly.

Co-Chair Nerenz: And basically the discussion --- SMP can just bullet page four. I just think there's some interesting things we can remind ourselves.

We basically have three measures that are very, very similar to each other, some are conceptually, some are data sources, same patient population, same entities being measured, same analytic methods, same developer, same, same, same, and if we look at what you see on page four which I think helps look at the relationship of all three of them.

3720 and 3718 got exactly the same ratings on reliability, but 3721 was different from the other two. On validity, 3721 and 3720 got essentially the same votes, but 3718 was different. So I think when we get into these discussions measure by measure, we should try to concentrate on what was it about that measure that made it different from other ones and as always we do, let's focus on the no pass and look very carefully at the developer responses, hear about that.

Let's look at the CNR. It may turn out that by the time we get to 3718, as Matt said earlier this morning, we may have nothing to discuss there. But it's there just in case our thinking on the first two shifts in any way our thinking about the third one.

So I just wanted to set this whole set up so we keep thinking about the relationship with each other.

Dr. Pickering: Great. Thank you, Dave.

Any questions from the SMP before we get started with our first measure?

Okay, all right. So seeing no hands raised, nothing in the chat, we'll proceed.

#3721 Patient-Reported Overall Physical Health Following Chemotherapy Among Adults with Breast Cancer (Purchaser Business Group on Health)

Our first measure in this set is 3721, as you can see on the screen which is the Patient-Reported Overall Physical Health Following Chemotherapy Among Adults With Breast Cancer. So this is a PRO-PM or patient reported outcome performance measure which assesses overall physical health among adult women with breast cancer entering survivorship after completion of chemotherapy administered or clear intent.

Overall, physical health is assessed using the PROMIS Global Health Version 1.2 scale administered at baseline so prior to chemotherapy. In that follow up which is about three months following completion of therapy, this measure is risk adjusted. It is a new measure. It is set at the clinician practice group level and like I said, the developer is Purchaser Business Group on Health.

Our lead discussant for this measure is Zhenqiuu.

You can find more of the assessment of reliability and validity in the discussion guide on page 13.

So this measure has a no pass on reliability and also a CNR on validity. So similar to how we've done this morning, the group will discuss these concerns, listen to any developer comments and responses and then move to vote on these issues as needed.

So we'll first start with reliability, so the reliability testing was conducted at the counter level and the accountable-entity level. So at that counter level, or the data elements level, reliability testing from the literature demonstrated that the PROMIS Global Health, the Cronbach's alphas are .92 for the overall, .81 for physical health, and .86 for mental health.

Related to the accountable-entity level testing, the test reliability measures for a signal to noise analysis was conducted and an estimate of the adjusted interclass correlation coefficient was .034, estimate of the reliability at the average sample size for group which is 32 patients per group was .534.

And then using the Spearman-Brown prophecy formula, the developer did estimate that in order to obtain a nominal reliability of .7, a minimum sample size of 66 patient respondents would be required. So the group's specific reliability ranged from .18 to .70 with a mean of .45 and a median reliability of .44.

The proportion of groups in a sample that has sufficient reliability using reliability threshold of .7 was 10 percent. So unlike some of the other measures, as Dave had mentioned in this group, there was significant concerns with the accountable-entity level and reliability testing results as only one of the ten groups involved in testing at reliability of .7.

So up for discussion are to access SMPs just to review the developer's responses to any of the SMP concerns and then re-vote as needed.

So Dave, with that summary, I'll turn it over to you for the reliability discussion.

Co-Chair Nerenz: And I'll immediately turn to Zhenqiu, nothing to add in the transition there.

Member Lin: Okay, so for this measure, the initial review, the vote was two moderate and six, low, so it didn't pass.

So the developer did respond to the comments from SMP. And the way I read it, the basically reiterate and measure entity level with a better result. I don't see new information pertaining to the reliability testing.

So my question is so given that there's no new information, you know, is presented, what are we going to do? Do we need to re-vote and the developer can correct me if I'm mistaken, right? So my question is for the Cycle 2, what do we do? I mean maybe respondent is more or less the same information.

Co-Chair Nerenz: Why don't we give Rachel an opportunity on this because in terms of what we have in front of us in writing, it did look like new information, but they absolutely missed something.

So Rachel, why don't you and your team have a chance here?

Ms. Brodie: Okay. I appreciate that. And thank you all for the sort of careful consideration that you've given all three of our measures.

We appreciate this opportunity to talk to you because we think these are important measures that there's a known gap in PRO-base measures in cancer care and particularly for patients with earlier stage curative use.

So in terms of the reliability testing, as noted, physical health or 3721 did not pass, but pain interference and fatigue did. We -- I'll go over some results because there is one certain new aspect of the testing, but what distinguished the other two measures is that their overall reliability of the performance measures was greater than both the .6 and the .7 reliability thresholds whereas physical health or 3721 was lower.

In all the reliability testing in our submission, we based that on the .7 reliability threshold, but we do want to note that SMP recommended acceptable thresholds is .6. So in our response, our developer response, we did add information about .6. And it does make one important point that when we use the .6 threshold, it does reduce for the minimal sample size needed to obtain that reliability to 43 patients.

In addition, when we used .6, 50 percent of the groups have reliability of .6 or greater. One of your comments did state that in using the .7 reliability only one of the groups reached that .7 reliability. So I did want to mention that was one piece of additional information.

Co-Chair Nerenz: Thank you. I see Patrick with a hand up.

Member Romano: Yes, I think -- I'd just say I appreciate the value of these measures and what they would bring to the overall measurement portfolio.

I'm curious, obviously, I think all of us, most of us picked up on the fact that of these three measures, this one was authoritatively and quantitatively different on reliability. So if you just left it to the adjusted ICC estimate, we talked about the interpretation of that estimate this morning, but it's three times higher for the other two measures, roughly .09 to 1 versus this measure .034.

The Spearman-Brown adjusted ICC estimate for this measure was clearly lower than for the other two and so I'm curious to hear your thoughts about that, if you've done some empirical analysis to better understand why this measure has lower reliability, one kind of answers the other. One concept that came to mind for me is that you referred to the Cronbach's alpha coefficient. Obviously, that is limited as a measure of internal consistency and reliability, but this measure is .81 whereas for the other two measures, it's .99 and .86. So this measure clearly has lower internal consistency reliability, although it's still above .8.

So there may be other factors that obviously all three of these measures have the same denominator. So it's not immediately apparent why this measure would be markedly worse in terms of its accountable-entity level reliability.

Ms. Brodie: And I'm not sure whether if Feifei or Kris, my colleagues would want to weigh in. But we do -- we do acknowledge that this measure has lower reliability. It requires a larger number of patients in order to get reliability. So we recognize that this measure needs more testing.

I want to over-speak to that. It's decent reliability for a PRO-PM, and there aren't really standards, but it is definitely lower.

Kris or Feifei, do either one of you want to address Patrick's -- Dr. Romano's comments in more detail?

Ms. Ye: This is Feifei from RAND. I think from empirically in terms of data, what we see is that the -- between like group variation for the physical health, just smaller than the -- I mean than the other two measures.

I think now we don't -- we haven't done analysis or like can figure out like, why, like the groups have lower between-group variability for this measure. But yeah, but that's just what we noticed from the scale.

Member Romano: Thank you, that's helpful.

Co-Chair Nerenz: Thanks. Paul, you have a hand up.

Member Kurlansky: Yeah, actually if I understand what Feifei was just saying, it sounds like the signal is smaller, so the signal-to-noise ratio may change. But I just, more of a practical question, in other words, for .7, you needed the 66 patients, for .6, you needed 40-some odd patients.

And the number of patients in each group that you actually tested was I think 32. So I was wondering if it might not make sense to just go back and test larger groups and see what your reliability actually is.

Ms. Brodie: Yeah, I mean, our intent is to continue to -- I mean, testing these measures. We would need to do additional testing in the maintenance phase, particularly on a measure that makes it to implementation. So you know, we're committed to continuing the test-based.

No, the average group size was 32. The minimum required sample size for .7 on the pain and fatigue measures were 22 and 23, fewer if we're looking at a .6 reliability threshold. But the physical health needed more patients to reach reliability.

Member Deutscher: Hello, this is -- this is Daniel. I'd like to just ask one question. Do you have a reason to believe that you will be able to have larger sample sizes for this measure with more data? Or is there something limiting the sample sizes for facilities, related specifically to this measure compared to the other two?

Ms. Brodie: It wouldn't be specific to this measure. I mean, we did test this measure during the pandemic, and that definitely affected our sample size. We did feel that we had enough data to fully test the measures.

We believe that when -- if any of these measures were implemented outside of the public health emergency and also in the context of a reporting program, that we would certainly have, you know, greater data.

Member Deutscher: Okay, thank you.

Member Lin: I do think that small testing sample size, I mean, does create some problem. As we were getting to the -- you know, get to validity, right. You have 323 respondents from ten groups, so it creates other issues as well.

Member Romano: Yeah, to that point, could you clarify, so you said in 2A05 that testing was planned on a sample of 21 oncology groups, but due to the impact of the public health emergency, only ten sites submitted sufficient data for inclusion in testing analyses.

So what was your threshold for sufficient data? Is there a specific numerical threshold that you implemented?

Ms. Brodie: We -- Feifei or Kris, please weigh in if I don't get this right, but we needed the test sites to have at least ten baseline and followup surveys, at least ten patients with baseline and followup surveys to include them in the testing sample.

So ten sites had that, and of course the numbers per site ranged significantly.

Member Romano: Yeah, I mean, it's just, you know, it's obvious, as others are saying here, that the solution to your problem is to collect more data and to be able to raise that threshold, that minimum threshold from ten.

You know, it's very common in measures to have minimum thresholds of 20 or 30, which would effectively resolve your problem. But we all realize, those of us who are developers, that Covid threw a wrench into all of the efforts.

Co-Chair Nerenz: I'm sorry, I was muted. Any other comments or questions from Subgroup 2 on the issue of reliability for this measure?

Member Deutscher: Maybe just a quick comment that probably leads to the discussion on the validity. From my perspective at least, the reliability problem flows into the validity issue in this case. And it's true that the differences are not -- are not very large between the three measures.

But when we look at meaningful differences between entities, we do find only one that was meaningfully different in this case compared to two for the other two measures.

So again, not a huge difference. But it looks like this is impacted by the reliability issue and could probably again be solved with larger -- larger samples in the future.

Co-Chair Nerenz: Well, I think just as an observation, it just, it may be that among these three measures, this one is just conceptually a little softer, a little broader, a little harder to get tight bands of measurement. And the other two just behave differently. It's just sort of the nature of the concept, possibly.

I saw Larry was your hand up? It came and went quickly on my screen.

Member Glance: Yeah, but I was going to speak when it's my turn to speak since I'm not in this subgroup.

Co-Chair Nerenz: Well, let's -- let's going once, going twice on Subgroup 2, anybody else Subgroup 2? Okay, now it is your turn.

Member Glance: Thanks. This is kind of a more general question that I wanted to bring up to the group. So we spent a lot of time talking about reliability thresholds, and then I think we came up with a tentative threshold for overall reliability.

And then we talked about the fact that we wanted to see reliability in different subgroups, meaning, you know, instead of just the median, maybe quartiles, by quartiles. But I don't think that we necessarily came up for thresholds for the quartiles.

So in other words, if we assume that an acceptable median reliability based on, say, splits -- say the usual signal to noise ratio is .6. Does that mean that at the low quartile, that the reliability has to be .6 as well?

Because by construction, if the reliability in that lower quartile of volume is .6, then the median reliability is going to be quite a bit higher than the threshold that we've set.

And I guess my point being is that I think it's very important that we're consistent across different measures in terms of what we ask as -- what we consider to be a reasonable threshold. If it's going to be .6 for the median, then I don't know, you know, we shouldn't be asking for .6 for the lower quartile.

And I just want to bring that up as a matter of discussion for the group.

Co-Chair Nerenz: Good point, fair enough. Any responses to that point specifically? I do think we have in front of us some median estimates of what I -- obviously a valid point, the math is clear. That if you've got a threshold for a median, you're going to have a lower expectation for a lower quartile.

And to my knowledge, we never tried to establish that. Or for an upper quartile for that matter. So, valid point. I'm not sure what we do with it.

Member Glance: But my point was that I think that the median reliability for this measure was what, it was over .5, .6, correct?

Co-Chair Nerenz: No.

Member Lin: It was .44.

Member Glance: Point 44, so it was right --okay.

Member Lin: I think mean is .45, median is .44, right?

Ms. Brodie: Yeah, mean was .45 for this measure.

Member Glance: Okay, thanks.

Ms. Brodie: For overall it was .53.

Co-Chair Nerenz: All right, how about -- well, let's just make sure there's no new question, comment, input. What's the pleasure of Subgroup 2? Is there any request to revote this? And again, you can speak up, you can chat, you can whatever you like. We'll give a little time here.

Dr. Pickering: If you want to direct message one of the team members, NQF staff, you could do that as well and remain anonymous.

Co-Chair Nerenz: So far we just have one request to not revote and nobody asking for a revote.

Dr. Pickering: Confirming here. Nothing on our end.

Co-Chair Nerenz: One more no revote. Okay, let us then move on to validity. ZQ, the ball's back in your court.

Member Lin: Okay, so I think SMP raised a number of questions, and then the developer did respond to each of them. And I'm just going to go over the key one.

So the first, develop assess face validity, right. They have that panel comprised of 12 members and eight vote -- or eight vote in favor of the -- of this measure. However, four members declined to vote due to concern about the very limited testing data and potential Covid impact.

So I think some, as some team members shared the same concern about unlimited testing data, right. And it does have material impact given that we only have 323 responded from ten sites. For examples, the developer used a modified Elixhauser comorbidity tool to identify comorbidity around this cohort.

And they were only, out of 26 potential comorbidity variable, they were only able to identify three with a sufficient response, right, they had very high potential in depression. And that's a direct impact of what would be included in the risk adjustment model, right. So it does create a issue about -- because this is a risk-adjusted measure.

If you look their updated research as a result, I mean, in the original submission I think there are some copy and paste issues because they have negative standard error. And also some read a larger error and the small P value.

But if you look at the updated table you can see that out of 13 risk variables, only three statistics they did count. And many can -- are not significant. So I think there's no sort of internal/external validation about a risk model. So you know, I do think there are legitimate concerns about the recent model as well due to the limited testing data set.

Additionally, a number of members raised the issues of high level of missing value, particularly for risk adjustment variable. For example, I think performance data, that baseline had 14% missing. And the other one's an aromatase inhibitor, like about 12% missing.

And it's not clear how those risk adjust variable, you know, missing with how they were handled in the model development and the measure score calculation. So I think because it's substantial.

And developer did respond, right, they anticipate once this measure is in a context of reporting, the value, the missing value, would be reduced. And also they think those data are not truly missing in the system, it's just not being captured.

But that's sort of typical, right. And in cardiovascular we know injection fraction, well, heart disease patients only available in the EMR. But we are still seeing high data value not captured. Just because they are not there, it doesn't mean they will be captured. So that's a missing value.

The other one is, a member also brought up about non-response, right. So developer respond by provide the response rate across -- they calculate respond rate in two ways. You do see a substantial variation across that.

And I also noted respond rate is related to respondents' marital data and insurance data, right. So how potential non-response bias should be handled, right. It's tricky.

And developer did point out that they tried to adjust marital data and insurance data. And it didn't seem to make it different, even contributing to model performance. But that's on top of a very small data set, right. It's not surprising you're not going see much different. Because even as is, you had 323, you had 13 variable. Ten are not significant to begin with, right.

And then developer also referred to a paper about, you know, once risk adjustment is conducted, non-response bias will go away. But that paper is for HCAHPS, I think. So it's different outcome, right. That's based on one way of a survey.

And for this particular measure, you base on -- need to have both based on baseline survey. And also three months after completion of chemo. So situation are somewhat different. I'm not sure the conclusion will automatically apply to this measure.

And then, so this is about potential non-response bias how it should, you know, it should be handled. And Daniel already mentioned, right, there's also a small meaningful different, only one of about ten group have a significant difference from the overall mean.

So that's one, two, three, four five. The last one is about I think is a question about timing of when baseline survey should be assessed, right. Because of -- this is a chemo patient, so there's two type of chemo. IV chemo, the baseline was down within the two week before IV chemo.

But for oral chemo, it's that two week before, also one week after. So there's potential -- I think there's a member brought up concern that if you do it after they started oral chemo, the side effect of the chemo may impact the baseline survey score, right.

So I think the developer's response is they want to include a capture of patient because of it's hard to ascertain the oral chemo start date. And also they don't think -- they think in general the side effects of chemo agents should not interfere with the baseline survey score.

So those are the concerns I captured from the review, and also some of the feedback from developer. And developer also provided additional result on validity, and they did provide additional empirical validity result. So that was added to report to the group, so.

Did I miss anything, David

Co-Chair Nerenz: Muted again, sorry. No, that was a quite excellent summary. Let me just add one thing, and I'll just speak to my own vote on this in the initial round, and perhaps we can take one set of things off the table for discussion, we'll see.

I was one of the two who voted insufficient on this. And it turns out my rationale for doing that was wrong, it was in error. When I was reading the original text, I was go through the validity section, there was reference to the TEP.

But I didn't see any listing of who the people were, what their credentials were, what their backgrounds were. And on that basis, pulled the trigger too quickly, I voted insufficient. Because I think we need to know who's on a technical expert panel voting face validity.

After I sent that in, I had occasion then to go scroll all the way down to the bottom of the document to the appendix, and sure enough, here's the list of the TEP members who, some I know personally, they're very well-qualified, good group.

And then in addition, we find that there's a face validity rating panel that now we know more about. So the basis for my vote just turns out to be wrong. And unless somebody else wants to stay on the issue of the composition of the face validity panel, personally I'm happy to just let that go, I am satisfied.

So we have a number of other issues on the table. Real quick, other Subgroup 2 members want to weigh in before we hear from the developer team?

Member Romano: This is Patrick. Yeah, I mean, this raised a number of questions about how we interpret face validity. Because obviously the evaluation here rests so much on face validity. And it is concerning when four members of the expert panel don't vote.

We appreciate the developer's honesty about explaining why they didn't vote, that it was not a random choice not to vote. Essentially, they didn't vote because they had concerns.

And so we need to interpret, you know, that as part of the assessment of the face validity that the face validity results were less optimistic than what the 8-0 vote suggests. So that's one issue that I think we've all been wrestling with.

We do appreciate the additional information about the patient and caregiver engagement, that that engagement was present through the process. Typically we see more diverse TEPs that include engagement from multiple stakeholders on the TEP itself that provides the final vote, but we appreciate that there was involvement through the -- through the process.

But otherwise, yeah, I'd be interested in hearing the developer's thoughts about the issues that ZQ has raised with respect to risk adjustment.

And Daniel Deutscher raised the issue that when we can only identify one group that qualifies as being different than the others on performance, and when that one group is worse, it suggests that the sample size may be inadequate, that there's a problem with the design of the measure in terms of the ability to identify better-performing groups.

And that's an issue because we like the idea of some symmetry that we'd like to be able to identify groups on both ends of the performance distribution.

Ms. Brodie: Yeah, and just maybe to start with that particular issue. There are several issues here to address, but I'll try to make sure I do so.

But to start with that particular issue, that was also one thing that distinguished this measure from the pain and fatigue, in that, in the case of the pain and fatigue measures, 3718 and 3720. They each had a group that was statistically different from the better performing and also one worse performing. Whereas this only had the one group that was statistically different.

I would say, though, that as I discussed in the reliability assessment that if we're using a .6 reliability threshold, 50% of the groups are statistically different.

One thing that we did add to the discussion around meaningful differences -- give me a second -- is we wanted to talk a bit about the P score. And what the literature tells us that meaningful important difference in cancer patients --.

So in the physical health measure, for example, the literature in the cancer population has suggested that a meaningful difference can be defined as between a 3 and 6 point difference on a P score scale, and the mean is 50, a standard deviation of 10.

So among the group scores that were significantly above or below the average, that mean absolute difference between the group score in the overall average was 5.19 points, which is more than half of the standard deviation: 5 points. So those results indicated that the PRO-PM measure could distinguish between groups' performance. That was our interpretation.

One other note on meaningful differences in the measure is that the -- when we look at the adjusted group score, which range for this measure from 40.34 to 43.49, with a standard deviation of 2.63, the confidence intervals for the highest through the lowest group scores did not overlap. But as you noted, one group had a significantly higher score than the average.

I go back to the other validity issues that were noted. Thank you for acknowledging the clarifications that we made around the patient engagement. There were two patients on the TEP, even though they served in other roles.

And we had two patients that were on the steering committee. And we also engaged a patient and caregiver council from the Michigan Oncology Quality Consortium at two different points to help us determine -- select which outcomes to measure and also to select a survey instrument. And those results were brought to the full TEP.

In terms of other empiric -- since the concerns were raised about the safe validity testing, we did want to provide the sort of initial empirical validity testing that we had at the measure level.

We went through a process with our technical expert panel where we asked them to rate the correlation that they would expect between publicly available and commonly available quality measures and the outcomes of our measures.

And there were four measures that we used because the TEP estimated that it would be -- rated these as sort of a moderate level of agreement. And then we ran the -- so we collected the data from those test sites but we didn't necessarily have those four measures for all ten test sites.

So we used the four measures that had at least seven test sites with the measures. So we did provide those results. And what you saw is that the correlations were in the moderate range, which agreed with what the TEP estimated or hypothesized and in the appropriate direction.

So an example would be if you're more likely to recommend the hospital or the care services or the degree to which the care was coordinated, that would be associated with lower pain, lower fatigue, or higher physical health.

We wanted to provide that additional information. It wasn't required as part of our initial submission, but we wanted to provide that.

In terms of risk adjustment, and having a small sample set to evaluate the risk variables, I don't -- I think that would be best for eight Feifei or Kris to help provide additional explanation.

Ms. Ye: This is Feifei from RAND. So I think the three issues that mentioned like regarding the risk adjustment model, missing data, non-response bias. So I will talk about these three.

And one common theme from this is kind of like the -- I mean, that has been repeated here -- is the small sample size, like the 320 over ten sites.

So first for missing data, so the missing data for the risk adjustors as well, those were taken care like by using the group mean imputation. So that it was like the -- how that was addressed in the modeling.

So that is like, for example, for comorbidity, even there is missing data or those subjects with missing data, they were imputed -- I mean, their values on morbidity were imputed. So to make sure like the, you know, the Hapeville model contained the complete sample of 320 -- the past 320 there.

And for non-response bias, we acknowledge that we like used like the case mix. I mean, we basically cited the CAHPS, like that paper. Which, I mean, the NHAMCS/NIS paper. That's the case mix assessments -- is more efficient. I mean, actually has like it's more efficient in addressing the non-response bias than the using response weights, especially for small sample size.

And we had like, if like as -- I mean, if it were going as planned, like, we have sufficient sample size. We actually would do the -- I mean the response -- non-response weighting and did that. And we'd based all the propensity score on message.

But that actually is limited by the current sample size, because propensity score weighting also requires large sample size. So that is why in the end we decided to just go with the case mix adjustment.

And one potential argument, not like that strong about it, but potential argument is that we didn't include the baseline data and baseline measure score. And these like, I mean, these variables like marital status or the insurance, like they tend to affect the baseline score as well.

So even though in our case mix adjustment model we test with these slight variables that kind of like related with non-response, we test with them and without them and see no significant difference. And so we decided not to include them just for parsimonious model purpose.

But if they are like related with like these scores with these like measure scores, we can say that they are partially adjusted by the baseline score, which was included in our model as the case mix adjuster.

So and also related to like, I mean, when Zhenqiu mentioned the comorbidity, for example, we tested all those like 23 like comorbidities and only selected just three. It's, again, because of the small sample size, we acknowledge that. And we had a lot of discussion over the risk adjustment model.

We do think that with, like, for the future testing our recommendation is to still consider all these like case mix adjusters we have tested. Because those were carefully selected from the expert panel review, from your know, like by the team. And also the expert panel.

So our recommendation is that like they should still be candidates for consideration in the case mix adjustment model like for future testing.

Co-Chair Nerenz: I don't want to cut anything off here. Paul, I see a hand up. Do you have something that's what directly related to what we just heard? If so, we should probably slide it in here.

Member Kurlansky: It's a simple question. The -- excuse me, the survey administrated for the associated measure that I was going to be reviewing, the administration rate was 84.5% and the response rate was 43.9%. So therefore, the overall response rate was 37%.

I'm just wondering, is that true also for this -- for this metric as well, that the overall response rate was 37%? Because at that rate, your ability to impute is just really very limited.

Member Lin: I think it depends on how you calculate that. I think it is -- it could be, I think that one way you calculate it 38 and the other way is about 40-something. So it may be slightly higher, but still in the same neighborhood.

Member Kurlansky: Same range.

Ms. Ye: Yes, and we -- I think this measure has been used, at least in the CAHPS work I have been involved with. So in the end CAHPS also has about like 30-40% response rates. And but their sample size is much bigger.

So I do, I mean, our data is limited. I mean, even with the imputation, still it's limited by the sample size.

Ms. Brodie: And one clarification if it helps. It is a single survey when it's administered where the respondent sees the PROMIS global questions. Then the pain questions, and then the fatigue question.

Member Lin: I actually had a question for one of our panel members. I see Dr. Walters is a medical oncologist, so I want to ask you about are you concerned about the timing of baseline survey score? Because it's a key adjuster, right, for the -- for the measure outcome.

If you look at the performance, that's the one biggest risk factor. I mean, I account for a majority of varying reduction. So for oral chemo, if patient started oral chemo, then you assess the baseline survey. Do you worry about the assessment may be impacted by the potential side effect of chemo agents?

Member Walters: So I am going to be careful what I say, but yes, the chemo in this circumstance was heavily, heavily, heavily predominant intravenous because it was breast cancer. If it was some other cancers, the percentage of oral chemotherapy might well be significantly enough to impact what you said.

But for the most part, in breast cancer, it is not.

Member Lin: So you don't anticipate many oral chemo breast cancer patients anyway, right?

Member Walters: Not for this population. Other populations, yes.

Member Lin: Okay, thank you.

Co-Chair Nerenz: All right, I'll confess to perhaps needing a reset on the sequencing. ZQ, you had six points, and I know there's been response to a couple of them. Should we flip back to Rachel and her team? Do you have more separate points to respond to, sort of, ZQ where are we in the flow of this now?

Member Lin: I don't know, I think we had a bunch of back and forth, right, so we, you know, I just summarized the feedback from the panel. I didn't -- Rachel did provide and Feifei both provide feedback. And they also articulate some additional information.

I guess it's well our subgroup, well, we are, what are we going to do?

Co-Chair Nerenz: Let me try one -- go ahead.

Ms. Ye: Sorry for interruption. Yeah, I think there's one thing I want to add like about the meaningful difference. Like as Rachel mentioned -- I mean Rachel mentions that for fatigue and pain it's kind of like one measure. I mean, there's one groups that can better and one groups with it's lower like than the average.

And for physical health, there is only one group's even better. But just by looking at the competence interval of like each size mean, there is another group actually is kind of marginally significantly lower. It just right touch the -- I mean comes into a right touch, the average there.

So I will say that there's a tendency, but again, given the issues we have seen for physical health measure, I mean, that is, yeah, that is not insignificant statistically at one point coupled with the other point.

Co-Chair Nerenz: All right, Daniel, hand up.

Member Deutscher: Yeah, I think thank you for this last comment. I think what we're seeing here overall, if I'm trying to think of an overarching theme for validity, is really just the issue of sample size.

I mean, obviously I also support David's view about face validity. I think even if it's not great, it has some issues maybe with some of the panel members not voting probably because of Covid, and. But I think the overall theme of that is fine. I would be fine with passing the face validity issue.

My main issue is really with the threats to validity, and mainly those meaningful differences of the ability to demonstrate that. And again, following this last comment, I think it's really just an issue of sample size.

Or at least there is a good chance that that's the case and this measure just needs more cases per side, and if possible more sides, which would make it look much, much better.

So this is at least my main issue for validity. I would like to see this measure retested basically with more data. Thank you.

Co-Chair Nerenz: Anyone else, Subgroup 2, questions, comments? Watching the clock a little bit. I think we're okay, but make sure we move along. Anyone else on the SMP, comment or question to make?

All right, speaking simply as a member of the subgroup, I think I would like the opportunity to revote just because of what I said awhile ago. I think my initial recorded vote was in error, at least the rationale for it was in error. And I'd like to have the chance to correct that.

And I know we've had some additional information come in front of us. So I would request a revote on this one, on validity.

Dr. Pickering: So that's enough to carry it over to a revote. So with that, if there's no other questions or comments from the SMP, we will go ahead and open up the revote for validity. It's the last calling for final questions and comments from the SMP.

Patrick.

Member Romano: Yeah, I think that obviously we're discussing 3721. We're considering the three measures as a group as they were submitted. And I think it's an interesting problem, because we have a problem here that the entire testing upon which reliability and validity was based was 323 patients from ten entities.

And we don't have any strict lower limit of what constitutes an adequate sample or testing a risk-adjusted outcome measure. But it's clear that this is in a gray zone. And it may be just sufficient to get us over the limit for the other two measures and perhaps fallen short on this measure.

But it's just a practical challenge that obviously all of us as developers has faced during the pandemic. I don't know necessarily what the solution is, but I'm personally still feeling that this sample is just not large enough to bring forward based on all the concerns that have been raised.

Co-Chair Nerenz: Just quick response, and as always, important, valid points. The rules of the game for the new measure only require the developers to bring forward face validity information. They're really not required to bring forward anything else. So I think we just need to keep that in mind.

Member Romano: But they also have to demonstrate the issue of identified threats to validity, so that's.

Co-Chair Nerenz: That's true, absolutely. The whole thing can be a package. I just want to make sure we're clear on sort of what's required. How these two get combined if they're both present with concerns, like that's -- that's where ultimately the -- each of our votes come in.

Member Deutscher: And David, just to clarify, the fact that, let's say we pass validity but we cannot pass it because of threats to validity, right. Even though the basic evidence for validity we think was good enough.

Co-Chair Nerenz: I have to defer to Matt and staff on that one. We're into the deep water, the fine print of the rules.

Dr. Pickering: Yeah, and I'll just step in here. So for a new measure for basically the acceptable form of validity testing, keeping in mind the measure type, which is -- this is a patient-reported outcome measure, performance measure, instrument-based measure.

So we have the data element validity with that patient encounter validity, and of course the score level.

So the encounter -- the encounter level validity also is required for this measure, regardless of if it's a new measure or maintenance, which the developer has provided but it doesn't seem the concerns are focused on the data element.

So we're not looking at the measure score. And so for a new measure, as David pointed out, face validity is the minimum acceptable validity for a new measure for the measure score. But there still needs to be consideration of the threats to validity.

And so that consideration needs to have empirical assessments of those threats in which some of things we've been discussing today.

So with that, just keeping all that in mind, that this is a Pro-PM. Validity testing should be done at encounter level and at the score level. Score level, you can -- since it's a new measure, you can have face validity, which has been part of the discussion.

All of which, there needs to be a consideration of the empirical assessment of the threat stability.

Member Kurlansky: Just to be very clear about this, if we feel that the face validity testing was adequate but the threats to validity were not adequate, what do we do?

Dr. Pickering: So that will then weigh into your decisionmaking and your votes. So if there's an inadequate assessment, whether it be it's insufficient to make your decision or the results of any of the threats to validity are low in your opinion, that would factor into your rating for the validity testing.

Member Kurlansky: Okay.

Dr. Pickering: Same thing if it was a measure that was not the measure's maintenance -- a maintenance measure -- and the risk adjustment wasn't adequate, not appropriate. It would factor into your overall assessment of validity rating.

Okay, any other final comments or questions before we move to revoting on validity? Okay, all right.

So Gabby, I'll turn it to you and you can open the vote for validity, validity voting for this measure.

Ms. Kyle-Lion: Sounds good, thanks, Matt. All right, just give me one moment to get the voting pulled up here. As a reminder, this is just for Subgroup 2 members. So if you're not in Subgroup 2, you should not vote on this -- on this measure.

So, voting is now open for Measure 3721 on validity. Your options are A for high, E for -- or sorry, I'm sorry. A for moderate, B for low, or C for insufficient. And I believe that there are ten Subgroup 2 members on the call, so we are looking for ten votes here.

Member Deutscher: And just a question for my understanding. Why isn't there an option for high for validity?

Ms. Kyle-Lion: There was no accountable entity level empirical testing done. So there was just face validity and patient encounter data element testing. And when no empirical accountable entity level validity testing is submitted, the highest option is high. Or sorry, the highest option is moderate.

Member Lin: Just I had the same question like Daniel. Given that developer did provide some supplementary information, right, they compared that. I'm saying how you should vote, I mean just acknowledging that they provided additional information associated score weight for outpatient Press Ganey and score and also HCAHPS score, so.

Dr. Pickering: So that's just supplemental information they have provided. It wasn't what was in the original submission. So right now it's just basically any testing that would be assessed here.

Member Deutscher: I still have a question about the face validity. Because the face validity question was on the entity level, right, the actual question that was asked.

Dr. Pickering: Right --

Member Deutscher: But that doesn't not account -- it was inferred, okay.

Dr. Pickering: Right.

Ms. Kyle-Lion: So we did get one vote via chat, so we are at ten votes now. And I will go ahead and include the vote that we got via chat in my final calculation here. Just give me one second to pull up these results.

So voting is now closed for NQF No. 3721 on validity. Okay, so including the chat vote that we received, there were two votes for moderate, five votes for low, and three votes for insufficient. Therefore the measure does not pass on validity.

I will pass it back to Dave and Matt.

Dr. Pickering: Dave, any final comments on that before we go to 3720?

Co-Chair Nerenz: No, no, I thank -- I thank everyone for careful attention to the various issues. We got a number of things in the air all at the same time, and I think thanks to our developers. And we'll move on, and we'll perhaps cover some of the same ground again.

#3720 Patient-Reported Fatigue Following Chemotherapy Among Adults with Breast Cancer (Purchaser Business Group on Health)

Dr. Pickering: Right, thank. So we'll go to 3720 because that also has a CNR on validity. This is obviously a similar measure as we've been discussing. It's the patient-reported fatigue following chemotherapy among adults with breast cancer.

It is a PRO-PM as well which assesses fatigue among adult women with breast cancer entering survivorship after completion of chemotherapy administered with curative intent. Fatigue is assessed by using the PROMIS fatigue 4a scale administered at baseline or prior to chemotherapy, and then at followup about three months following completion of chemotherapy.

The measure is risk-adjusted. It's a new measure. It's at the clinician group practice level. Developer again is Purchaser Business Group on Health. Our lead discussant is Paul Kurlansky, excuse me.

Member Kurlansky: Kurlansky.

Dr. Pickering: Yeah, thank you. And then at discussion guide page 17 is where you'll find all this information. For that validity assessment, this is a data element validity that was conducted as well as score level.

For data element, the first -- the percentage agreement by data element ranged from 71.3 -- 71.63 to 100%. The reported cappers ranged from .64 to .67. The reported sensitivity ranged from 33.33 to 89.52%. And the specificity ranged from 60 to 99.45%. The data can be found in the table 2(b)(1). And several cells in this table were intentionally left blank.

The score level, the validity testing was -- major score was conducted. There was systematic assessment of face validity, as we've been discussing, using a panel of 12 oncologists. Eight of the 12 participated in the survey. All eight indicated moderate agreement, agreement, or strong agreement to the -- to the survey response.

For physical health, all eight agreed or strongly agreed that the measure did differentiate good versus poor quality. And four oncologists declined to participate in face validity voting, expressed concern regarding the impact of Covid on sample and those performance scores.

There were four exclusions: patients on an interventional and therapeutic clinical trial, patient who experienced relapse or disease progression, patients who have -- patients who leave the practice, or patients who die, all of which weren't in a lot of major concerns related to those exclusions from the SMP.

This measure is risk-adjusted, including 13 risk factors. The model discrimination was tested during the clinical -- during clinical trial. Comparing scores between null and the multivariate model adjustments for pain and adherence resulted in a value of .87.

It appears some correlation coefficient between the observed and the predicted responses was .55. For meaningful differences in performance, the mean group performance score was 48.51, with a standard deviation of 3.13, with a median score of 48.67 in a range of 42.13 to 52.07.

So two of the ten groups had significantly different scores in the overall average, one more favorable and the other less favorable. Among those two groups, the mean absolute difference between them overall average was 4.9 points on a T score scale.

And finally for missingness, the missingness ranged .00 to 093% for the PROMIS item scales. So similarly with some of the discussions we've had, there were some concerns related to the face validity testing, some concerns with some missingness as well. And also meaningful differences in performance.

So today we're just again discussing any of these concerns and see if we can revote on validity.

So Dave, I'll turn it back to you and see if Paul wants to take us through his concerns about this.

Co-Chair Nerenz: Let's go straight to Paul, nothing to add.

Member Kurlansky: As usual, Matt's done all the hard work for me, and actually ZQ as well because of the similarity of the measures.

I think it's reasonable pretty much to everyone in the committee or nearly everyone in the committee was reasonably satisfied with the reliability of the measure. So I don't think we need to focus on that.

And the validity side, there were two or maybe three major concerns. One we've already addressed really in the other measure, and that the term's the missingness of the data -- the missingness of the data, the missingness of the response or the limitedness of the response.

But the other issue here is one of face validity and it's something that -- it resonates a little bit with what Patrick was saying but in this case it's perhaps even a little more compelling.

And that is, you know, I'm a cardiac surgeon, not an oncologist. But the fact that people should be fatigued after three months of metabolic assault does not necessarily -- it's very meaningful and it's very clinically meaningful. But that it represents variations in quality is not exactly clear to me.

And so therefore to me, the opinion of the expert panel in terms of face validity becomes very important. And so therefore, the fact that four out of the 12 did not respond or did not feel that they could respond because of the pandemic is -- takes on a little bit more meaning.

And then if you look a little more carefully, you know, the agreement was based on a 1 through 5 scale, but 2 and 3 is moderately agree and 4 and 5 is agree, agree or strongly agree. And actually, only three out of the eight who responded strongly agreed.

So to me, you have three out of 12 on the expert panel really strongly endorse or even agree that this is a metric of quality. To me, that is a major concern.

And it really questions, you know, for a non-oncologist, it certainly raises a question in my mind as to whether or not the metric itself really passes face validity, which is the core of what needs to be passed for validity.

I'm much less concerned about meaningful differences. There were two out of the ten, one higher, one lower. That kind of a distribution is actually not uncommon for other measures that we know to be meaningful and have high reliability.

So you know, it seems even with the limited sample size that there is an ability to distinguish. But whether or not that distinction it really represents an issue of quality is really I think the major concern.

The other major concern obviously is the one we've already discussed, which is just a the low response rate and whether 37 or 42% of responses is sufficient in order to really judge this measure.

And those are really the -- so I throw it open to my fellow group members for their comments and then obviously to the developers as to their responses.

Co-Chair Nerenz: Thanks, Paul, a really nice summary, you and Matt both. Daniel, hand up.

Member Deutscher: Yeah, thanks for these summaries. I'd like just to comment on the face validity issue. And it's true that in this case I agree it's a little bit more trickier than the other two measures because of those only three of eight expert panel members that voted I think it was agree or strongly agree.

But if we take it just one category, one response category lower, which is moderate or higher, then we have eight out of eight, exactly as we do for the other measures.

So I don't know that we have a strict rule about, you know, how to evaluate face validity and those ratings. I would relate I guess stronger to the concern that Paul raised if I would see at least one member rating lower than moderate.

But in this case, and especially considering how we rated other measures in the past regarding face validity, for me it's difficult not to pass the measure for validity just because of this. And since we have less threats to validity for this measure, overall I do have less concerns. Thanks.

Member Kurlansky: I should have pointed out actually that one of the reasons, in addition to the four that didn't feel they could respond because of the pandemic, the ones who were -- could not strongly agree also could not strongly agree because of pandemic. They felt that the pandemic was confusing or confounding the issue of fatigue.

So and that was distinct in this metric as opposed to the other two. Which makes me think that we just got to, you know, they just got to go back and reassess now that we are in a different phase of the pandemic.

Co-Chair Nerenz: Thanks, Paul and Daniel. Others in Subgroup 2 before we move to Rachel and her team?

Okay, I see no hands up. Rachel, it's yours.

Ms. Brodie: Thank you. I wanted to start by trying to address Paul's first comments around sort of the rationale for the measure and also the concern that, the fatigue being more sensitive to the pandemic.

So, first of all, the rationale for the measure is that patients who undergo chemotherapy with curative intent tend to have persistent symptoms and detriments after the treatment, including pain, fatigue, and health-related quality of life, including physical health, and those symptoms and outcomes of the treatment tend to persist for months and even years.

So, that's why the measures were important and that was related to the timing of assessing the numerator. And there is evidence that practices can, if they manage these symptoms, that they can position their patients better for entering the survivorship phase.

So, related to the concern around the pandemic affecting, the fatigue being more sensitive to the pandemic, one thing I think is important to mention is that we are assessing fatigue at baseline as well and that's used as a risk assessor, so I think that that's important to consider.

Another thing I would say, which I was going to say, but Daniel brought it up, is that while we had a face validity panel, only three votes of fours and fives was strongly agree and agree.

We didn't have any votes, any one or two votes disagree. The other three were moderately agree because they were like, well, this can be more influenced by what's happening in our country right now.

So, those are two things I wanted to raise related to face validity and the questions that came up. As far phase two, was there a question on meaningful differences? Sorry.

Member Kurlansky: Rachel, I didn't think there was -- I mean, there may have been from other panel members, but in my review, I didn't think so.

Ms. Brodie: Okay, thank you. And then Feifei or Kris, does anyone want to add to my response?

Ms. McNiff Landrum: This is Kris. I would just add that we cited the NCCN guidelines for cancer-related fatigue and these are among guidelines that have recommendations at the 2B level or higher, including a variety of randomized clinical trials that look at interventions that tend to be specifically focused on exercise or psychological interventions for cancer fatigue reduction.

And there is, for instance, a meta-analysis of 113 studies, more than 11,000 patients, that did show that exercise and the psychological interventions do significantly improve cancer-related fatigue for patients on active treatment, and many of these were conducted among breast cancer patients.

So, there is -- we know that it's under-assessed, under-treated, a significant issue that patients have if you ask them what bothers them most, and there are interventions that improve cancer-related fatigue.

Member Kurlansky: I was just wondering, did you happen to repoll your, now that the pandemic is in a different phase some would say, did you happen to repoll your panel?

Ms. Brodie: No, we -- the face validity panel was this year in April, March, April.

Co-Chair Nerenz: Okay, let me -- just a quick observation. I'm not hearing the same concerns about threats to validity here that we had on the first measure.

Is that because people just don't want to repeat themselves or is that because this one has fewer threats to validity? Any quick guidance on that? I just want to make sure we don't move too quickly to closure here.

Member Kurlansky: No, you know, in terms of the low response rate and those threats, I think, and sample size threats, I think they're identical. I just didn't want to go through the same discussion again.

Co-Chair Nerenz: Yeah, yeah, okay.

Member Deutscher: I think the one that's not identical is the issue of meaningful differences. We see more meaningful differences in this case than appear through the previous measure.

Co-Chair Nerenz: Okay, thank you.

Member Lin: Yeah, I think you can tell by the larger between variance, right? So, you could detect a larger variance, but in terms of simple side, the risk model and non-response, I think they are about the same. They're kind of using the same data set.

Co-Chair Nerenz: All right, so let us first just open for Subgroup 2. Any additional thoughts, comments? Our developer team, anything else you think we need to know? Any other SMP members who haven't spoken up yet, anything else?

All right, I'll -- I sort of feel the same issue on revoting I had on the first one. For me personally, it's the same reason. I don't know that that speaks to the heart of the matter, but I think if we did it on the first one, I'm feeling some sense of appropriateness to redo it on this one given the things we've discussed. I'd like the chance myself actually to do that. Patrick?

Member Romano: I'm wondering if Ron or someone could address his comment in the chat just to better understand what this intersection is between COVID and response rates and fatigue. So, is it the case that COVID fatigue is a threat to validity? Can we flesh out exactly what this burden is?

Member Walters: What I meant with that comment was how on earth do you distinguish those two? I mean, the developer said they did baseline values, but the baseline values happened to occur a lot of times before COVID really hit its strength too.

So, I don't know how those two are related in this particular analysis and I doubt very much there's any way possible to tease them out. That's just what that comment was about.

Member Kurlansky: Unless you now create a new risk model with COVID positivity or history of COVID in the model.

Member Walters: Yeah.

Member Romano: I mean, it's just obvious, of course, that many of us who have suffered COVID have experienced COVID-related fatigue and the symptoms are known to continue for a while.

So, and it seems like the oncologists thought this was a particular concern for this measure. So, it's just helpful to understand the timing of the baseline survey. So, do we think that a fair number of patients could have contracted COVID between the baseline survey and the follow-up survey and that is less likely to be a problem for the measure going forward?

Member Walters: I sure hope so. I wish we all knew the answer to that question. So, for those of you that didn't see it, my comment was is COVID fatigue a threat to validity? And that was kind of a rhetorical question, but it's one that I don't think we'll ever be able to sort out really.

I mean, is fatigue important? Does it happen? Absolutely. Did we happen to have COVID at the same time during a lot of the data collection? Absolutely. So, I mean, what do you do?

Member Kurlansky: Well, the reason why it's important is because unlike in the other two measures, even amongst the eight that did vote, I think four of them raised this issue. The four of those that did moderately agree raised this question as to whether or not COVID was confounding, which was keeping them from being more strongly agree.

Member Walters: You're right.

Co-Chair Nerenz: Matt, Gabby, unless someone else has an additional comment.

Member Romano: Do the developers have anything to add on this question?

Ms. Brodie: I would just add that, you know, when you have a patient who is recovering from chemotherapy who may or may not have had COVID, that there are interventions, as Kris mentioned, either psychological or exercise, et cetera.

And if you're looking at these patients and trying to give them the highest quality care and help them to enter survivorship in the best possible state, you need to be looking at these things and trying to address them. So, there is some actionableness, and we recognize that COVID on top of chemotherapy, that's rough.

And I would say that while there was some, a little more hesitance on this with the three votes, that there were not any disagree. They, you know, they were -- they thought it was an important measure.

Dr. Pickering: No other comments from SMP this last call? Thank you, Ron, for proposing to revote. All right, so we will go ahead and move to vote on this measure now, so I'll turn it over to Gabby.

Ms. Brodie: One more thing.

Dr. Pickering: Oh, yes?

Ms. Brodie: Not to be flippant, as high is not an option, but we did provide the additional empirical reliability, I'm sorry, validity measure results. I would hope that maybe that might influence some more moderate votes.

Dr. Pickering: Thanks, Rachel. Okay, all right, so no other comments and I'm seeing no hands raised, and thanks, Rachel, for the added plug there. We'll go ahead and move to votes on this measure, so voting on validity. This is just for Subgroup number 2, so I'll turn it over to Gabby.

Ms. Kyle-Lion: Thanks, Matt. Voting is now open for NQF number 3720. Your options are A for moderate, B for low, or C for insufficient.

Co-Chair Nerenz: I'm sorry, Gabby. Before we -- I'd like to get back to this issue of the high. We kind of shut that off the first time because sort of the theory is if the information hadn't come in the initial submission, then it basically didn't count, and I just want to challenge that quickly.

If the developers come back in a response, in all other senses, we take it into account, we talk about it, we use it to think about it. Why doesn't it count for this purpose?

Dr. Pickering: So, this is just because it wasn't submitted initially when the measure was first submitted to SMP. If it comes through in the responses, it can be considered in the discussion, but the measure is voted on as it's submitted initially. Does that help, Dave?

Co-Chair Nerenz: Well, it does. I mean, you folks are more expert on the rules. It just was -- I was feeling a little uncomfortable following on Rachel's comment that we now had in front of us some empirical validity testing and we're allowed to talk about it, but we're not allowed to vote on it, but if those are the rules, those are the rules, okay.

Member Romano: And I think perhaps we would be allowed to -- I mean, I guess that because the distinction between high and moderate doesn't really matter in terms of decision making, it could influence the vote from low to moderate, for example. Is that possible?

Co-Chair Nerenz: Maybe, but if the rules are that this is what we're allowed to do, but I guess, yes, I mean, to reassure the developers, in practical matters, the difference between high and moderate is essentially nothing, and the folks voting just have to understand that moderate is basically a pass and low and insufficient are basically not pass, and vote accordingly.

Dr. Pickering: And we just keep to how the measure has been submitted and left, so there may be some additional discussion that happens based on some analyses developers do, even as it goes through the standing committee, but we are looking at how the measure has been submitted to NQF because that is what --

The testing information and all of that, what the standing committee would also have available to them is what was submitted, and there could be some still discussion at the standing committee level on additional analyses the developer did take.

So, that's why we're still keeping to this moderate level or ceiling is because face validity was the testing that was done for the outlines global testing, but great discussions and thanks, Dave.

Okay, so we are -- I think we have all of the votes, Gabby. I'll turn it back to you.

Ms. Kyle-Lion: We're at nine votes. I was just going to give it one more second to see if the last -- we are expecting ten, so if anybody is having trouble voting -- oh, we just hit ten, sorry, never mind.

Okay, voting is now closed on NQF number 3721 on validity. Just give me one second to pull up the results.

(Pause.)

Ms. Kyle-Lion: Okay, so we received six votes for moderate, three votes for low, and one vote for insufficient. Therefore, the vote remains consensus not reached on validity. I'll pass it back to Matt and Dave.

Dr. Pickering: Okay, great. I'll just make a note that CNR decisions from the SMP, as a reminder, they do go to the standing committee as well, so they can also make an assessment here.

And related to the other validity votes, which was not passed for some of the other concerns we had talked about, you know, there is still an opportunity for the standing committees to pull the measure for discussion and we'll still determine whether or not it could be eligible for a revote as well.

But I want to thank the subgroup for evaluating that measure as well. Dave, I'll turn it to you to see if there's any need from the SMP to move forward with 3718 or not given some of the concerns that we've discussed. It may not be as much of an issue.

#3718 Patient-Reported Pain Interference Following Chemotherapy Among Adults with Breast Cancer (Purchaser Business Group on Health)

Co-Chair Nerenz: I'll do exactly that and I just need a sense of the subgroup given where we spent the last hour and a half or so. Is there anything that you've heard or seen on 3720 and 21 that would likely change your initial vote on 3718?

It passed both, particularly on reliability. It's like the last measure we just passed straight through. So, I think I'll turn first to our discussion leader and then to anybody else.

The form of my question is does anybody in Subgroup 2 feel the need to discuss and/or revote 3718?

Member Deutscher: Well, I'll just comment on that. From a group perspective, maybe there is because 3720 stayed at consensus not reached, so we'd have to identify the differences between 3720 and 3718, which did pass on validity.

From my personal perspective, no, I actually also passed the 3720 on validity and I don't have any new issues to discuss compared to what we've discussed up to now on all of the other issues and threats to validity.

From my perspective, there is no need to revote, but I think I'll pass it onto the other members of the group so those that do have concerns could bring them up here.

Co-Chair Nerenz: Ron?

Member Walters: Well, I just have the same consideration that was raised in all of our email discussions too as far as -- especially given the last vote because I agree that from -- we've separated out 3721 and that had problems that we clearly identified.

But this is going to be nitpicking, but the previous one then was revoted at 6-4 and this one was previously 7-3, so how do you explain those and does it matter anyway, except at 6-4, as was stated, it gets consensus not reached and 7-3 gets passed?

So, there's -- someone -- we may be asked by someone what was the difference between the last measure and this measure from a validity perspective and I know there's a lot of discussion about that that went on previously.

Co-Chair Nerenz: And apparently just on the math, one of us, one out of ten, felt a little more favorably about this one than the other one. Then the question is do we want to get into it on that basis?

Member Walters: It's like that oncologist in the 12 that was a three instead of a four or five.

Co-Chair Nerenz: Yeah, well, and I might turn to Matt. In terms of the progress of this forward, is it going to make any difference? You know, if based on what we've heard the last hour and a half, somebody who was favorable on this is now distinctly unfavorable for reasons that could be articulated, I could see why we really ought to work that through, but that's essentially what I'm calling out. I'm not going to lose too much sleep tonight over a 7-3 versus 6-4 initial vote.

Dr. Pickering: Yeah, I mean, the standing committees will probably look at this and want to know why this would pass and the other one is CNR. Well, I think the 3721 can then speak for itself, but they may want to know why this one passed and the other one was CNR.

So, but, I mean, both will be discussed, so there may be some teasing out that the standing committees can do. So, if the SMP is comfortable with where the validity testing is for 3718 and there's no need to revote, we can move that forward.

I would just note that, you know, we can carry over some recommendations from the SMP that, you know, the standing committee needs to carefully assess the validity across all two or three if they pull that just for consistency in evaluation.

But, you know, if no SMP members really want to, you know, propose a revote on 3718, you know, the votes can stand and we can just carry over that recommendation from the SMP about really, you know, considering the validity assessment of 3718 given light, you know, what happened with 3720.

Co-Chair Nerenz: Patrick, hand up? You're muted, Patrick.

Member Romano: Sorry, I'll plead guilty to raising this issue in internal discussion, you know, by email before this, and, you know, again, when I did a side-by-side comparison of 3718 and 3720, I couldn't see any convincing reason why 3720 would be rated lower on validity than 3718. So, that's why I proposed to have this discussion.

I am on the majority side of supporting a pass for both of the measures. So, I'm not in a position to say well, let's revote 3718 to lower its vote because I voted to raise the vote for 3720, but it does seem inconsistent that we've said CNR for 3720 but pass for 3718.

So, someone among us must believe that there's stronger evidence of validity for 3718 than for 3720 and I'd love to hear what that argument is.

Co-Chair Nerenz: Yeah, let me quickly respond to that and then turn to Daniel. I think we also have to be humble about the reliability of our own processes.

If we're talking about a 7-3 versus a 6-4 vote that went in the same general direction, you know, maybe somebody had a little attack of heartburn at the moment of voting and, you know, we just have to decide which of our differences across measures are meaningful and which aren't, and I think that's what's in front of us right now. Yeah, Daniel?

Member Deutscher: Yeah, so as Patrick, I'm also uncomfortable with the differing results for these two measures, but if I look again at the side-by-side comparison, the difference is probably those only three out of eight that voted agree or strongly agree on face validity. That would be my guess.

So, we do have eight out of eight members -- and again, I'm just guessing, so whoever made that call could speak for that, but if that is the case -- and as

I mentioned before, we don't have strict rules about that, right, about how do we rate results from face validity.

So, if eight out of eight agreed moderately or more and that is the reason for the different results of the votes from these two measures, is that a good enough reason?

And this may be more appropriate for an internal group discussion in the future if it's something we cannot change right now, and when I say change, I mean specifically for 3720.

So, then, you know, that's what we should do. We should discuss this later on and try to improve, as you said, David, the reliability of our own process regarding face validity.

And again, this is just, you know, this is the only difference I see between the two measures that could explain the differing results.

Member Kurlansky: I suspect I'm the guilty party and Daniel has identified exactly, I think, what the concern was. In other words, if you have eight out of 12 and eight out of the eight agreed, it's a different story than if you have eight out of 12 and only three out of the eight strongly agree and the other five, four or five are raising concerns about COVID that weren't raised in the other metrics.

So, it's a different assessment by the experts on face validity, which is the core metric that we used to judge validity in this measure.

Co-Chair Nerenz: Yeah, and that would certainly then be something that could be passed onto the standing committee about why one is pass and one is CNR.

Dr. Pickering: Agreed. I think that would be good for the standing committee to know from the SMP what the difference is. I think that's very helpful. Thank you.

And Ron and Patrick, thank you for your comment. Yes, it's great that we're having this consideration and discussion. It allows us to be consistent and also provide this to the standing committee so they understand how the SMP was differing in their votes.

But Dave, anything else we want to get from the SMP related to this?

Co-Chair Nerenz: No, I think we're ready to move on then and I thank everyone for a respectful, diligent discussion. I thank our developers for their input and response. And I do, again, want to reassure the developers that anything other than a pure pass here does not mean the measure is dead. It can still move onto the standing committee and they just take our discussions into account.

Dr. Pickering: Excellent. Okay, so just confirming, we will not be revoting on 3718. We did discuss the differences there, which we will include that not only in the summary, but also it will be relayed to the standing committee for their consideration as well.

So, I thank you, the SMP, for that time to talk through this, and also thank you again to the developer. As David had mentioned, we're very appreciative of your time and just, you know, to answer all of the SMP concerns and questions.

Okay, if there's nothing else related to measure evaluations, that will conclude our evaluation proceedings today. We have a few more items just to go through as usual, but I'll just double-check once more. Any other questions or comments from the SMP before we proceed?

Thank you, Ron. It looks like you also mention that COVID pain is less likely to be a threat to validity than COVID fatigue. I appreciate the comments in recognizing that.

Okay, we will go ahead and move forward just to wrap up our proceedings today. So, at this time, as always, we're going to see if there's anyone from the public who would like to speak up and provide any comments, if there's any comments for the group's consideration.

You can do so by just raising your hand on the application and, you know, it will get you to the top, and we'll call your name. If you aren't able to raise your hand, you can press *6 and we will just take your name and we'll put you in line.

NQF Member and Public Comment

So, first, I see Don Casey. Don, would you like to provide any comments?

Dr. Casey: Yes, Matt, thanks. Can you hear me?

Dr. Pickering: Yes, we can.

Dr. Casey: Great, thank you for the opportunity to provide public comment. I'm an associate professor of internal medicine at Rush in Chicago. I'm speaking for myself as a member of the Standing Patient Experience and Functioning Committee, and I thank the Scientific Methods Panel for their deliberations on 3721, 3720, and 3718.

I have one generic comment which would be related to, I believe, Dr. Romano's initial note to staff that it's been over a year and a half now since we've been cogitating on the excellent deliberations of the SMP regarding validity and reliability.

And I understand that this is still a work a progress into '23, it just can't happen fast enough for us to get clarified about that issue, understanding that it fits in with a larger scale issue of the overall consensus development process.

I have a specific comment. Unfortunately, the discussion guide that I was able to get didn't have information on 3718, but I assume it's similar, and I also don't have background, but I wanted to bring an important issue to the table here relative to scientific acceptability.

With respect to pages 73 and 84, which under issue three, provide a reference to a study done in the Journal of Clinical Oncology, which, as it turns out, was an abstract poster presented at ASCO a couple of years ago funded by Merck, and consisted only of 198 patients with breast cancer through a quality improvement research, non-randomized study related to the use of an app at six and 12 months of onset of the study in one center.

So, it was limited and it just raises my continued concern about being sure that we ask the measure developers to do a better job of providing evidence and evaluating it.

You know, in particular, promise ten contains both pain and fatigue as I note you know, but in this study for breast cancer, the only predictor of a worse promised physical health score was cancer stage four versus one for that particular subset because it did include gynecologic cancers.

And so, you know, having a mother who died of ovarian cancer, I can say that there may be causal interactions between the symptoms of fatigue, pain, and probably depression related to having separate measures.

So, understanding how the vote went, I just wanted to raise that set of issues for the panel to think about and be more cautious in thinking about the evidence that's provided in the future. Thank you.

Dr. Pickering: Thank you so much, Don, appreciate the comments. And Don, I know the version that we have does have 3718 and we'll have to make sure we confirm that is available online, but that 3718 does begin on page 21 of that discussion guide, but --

Dr. Casey: Okay.

Dr. Pickering: -- we'll confirm that.

Dr. Casey: All right, thank you. I'll look again. I got it online.

Dr. Pickering: Okay, thank you. Any other comments from members of the public? You can raise your hand, or if you're calling on the phone and not able to raise your hand, you can press *6 and unmute yourself.

One more time, if any members of the public wish to provide comments to the SMP, now is your opportunity to do so. I'll just do one last call. Thanks, Don.

Okay, all right, so we'll keep going. So, also we just wanted to recognize and thank a few of our SMP members who will not be continuing on going into next year, so their terms are ending at the end of this year.

You can see their names listed here and I will go ahead and announce them. It's Eric Weinhandl, John Bott, Joe Hyder, Joe Kunisch, and Terri Warholak, some of which are not on the call today, but we do want to thank you, all of these members listed here, for your service.

It is very much a valuable part of our work and really contributing a lot of your time, a lot of review time and getting on these calls, and really insightful thinking and decision making that makes this process work so well, and we very much appreciate your engagement and involvement here.

I'll pause to see if, you know, if anybody has any remarks or if Dave or Christie want to share anything for these individuals that won't be continuing on next year.

Co-Chair Nerenz: Well, certainly let me just add to the thanks. And I noticed in the chat, Gene is pointing out that his name also should be added there, so Gene, Eric, John, Joe, and Joe, and Terri, thank you so much.

To me in the role that I play in this group, it is a tremendous pleasure every time we meet to hear people's thoughts, to learn, to go through what is always a respectful and collegial discussion on some of these tricky issues and I just appreciate it so much.

And, you know, I think folks in the general public, for instance, the developers, see what happens live, so to speak, in sessions like today.

What they don't see are the hours and hours and hours that go on behind the scenes with reviewing submitted materials, perhaps exchanging notes back and forth, digging up references, trying to make sure that the votes we make are the best ones we can possibly make.

All of you who are rotating off have done this now for quite a while, contributed greatly to the work of NQF, contributed by extension to the whole quality measurement field. Thank you.

Co-Chair Teigland: I can't add much to that, Dave. Thank you for summarizing that so well. Yes, we hate to see these folks go. I assume we'll be rolling some new folks on, Matt, which always changes dynamics a bit, but great discussion.

As I said, we had some challenging issues to work through today. I think it gives us some content to talk about at some future off-cycle SMP meetings as to how we can make this process a little bit better, a little bit tighter, a little bit more transparent and consistent.

We're always working to improve as Dave said. We're learning from each other all the time, and thanks, everyone, for a great discussion today.

Dr. Pickering: Yeah, and thanks, Gene. I'm sorry I missed your name on there. I know that you're coming back from going overseas. I believe you had let us know earlier that you would also be coming off of the SMP, so thank you as well for your service and sorry your name was not on here.

Well, thank you very much, everyone. We'll go to this next slide and we'll just talk about next steps, so I'll turn it over to the team lead to cover the next steps.

Next Steps

Ms. Ingber: Thanks, Matt. So, as folks know, the full measure submission is for the remaining criteria and the deadlines for those full submissions depend on the topic area, so those dates will be November 1, 8, and 15. You can find more information about that on the NQF website.

NQF staff will be summarizing the relevant measure information and discussions of the SMP from today and providing those to the various standing committees who will evaluate measures in February 2023. The next intent to submit deadline will be on January 5 as well. Next slide?

If you have any questions or comments for our team, we're happy to receive them at methodspanel@qualityforum.org. Please don't hesitate to reach out. You can also use the phone number. The project page for methods panel related activities is linked there as well as the SharePoint site for our panel members. Next slide?

Are there any questions about the next steps?

Member Kurlansky: I don't have a question. I just wanted, while we're thanking panel members, which is totally appropriate, I wanted to thank the staff.

I've had the honor of serving on this committee for a few years now and the process was good before, but it's better now, and I think that the amount of work that the staff does to facilitate our work is really tremendous and it's deeply appreciated.

Ms. Ingber: Thank you.

Dr. Pickering: Thank you so much. It's great efforts.

We've always enjoyed these meetings and working with this group. You all are so insightful and just it's great to see great minds come together, and some of the things that we've been able to do together have been quite successful.

And we'll be carrying that over into next year as we started out that conversation when Patrick asked that question early on this morning, so we'll be continuing on into next year with a pipeline of things that we'll be looking at with evaluation updates with our criteria, so thank you for those comments.

Ms. Ingber: Yes, it's an absolute pleasure. Any other questions? Okay, I'll pass it to you, Matt. Thanks, everyone.

Dr. Pickering: Thank you. Christie or Dave, any closing remarks?

Co-Chair Teigland: I'm good.

Dr. Pickering: All right, well --

Co-Chair Teigland: I'll give back a little bit of time here. That's always a good thing. We're very efficient today even though we had some tricky issues, so thanks to everyone for moving it along.

Dr. Pickering: All right, thank you all very much. Have a great rest of your day and a great remainder of your week, and we'll be in touch via email. And as always, you know how to reach us, and have a good rest of your day. Thank you all very much.

Adjourn

(Whereupon, the above-entitled matter went off the record at 2:53 p.m.)