National Quality Forum

Scientific Methods Panel

Fall 2021 Measure Evaluation Meeting

Tuesday

October 26, 2021

The Scientific Methods Panel met via Videoconference, at 11:00 a.m. EDT, David Nerenz and Christie Teigland, Co-chairs, presiding.

Present:

David Nerenz, PhD, Co-Chair
Christie Teigland, PhD, Co-Chair
J. Matt Austin, PhD, Johns Hopkins Armstrong Institute for Patient Safety and Quality
John Bott, MBA, MSSW, the Alliance; the Leapfrog Group
Daniel Deutscher, PhD, MSCPT, Maccabi Healthcare Services
Lacy Fabian, PhD, the Mitre Corporation
Marybeth Farquhar, PhD, MSN, RN, American Urological Association
Jeffrey Geppert, EDM, JD, Battelle Memorial Institute
Laurent Glance, MD, University of Rochester School of Medicine and Dentistry
Sherrie Kaplan, PhD, MPH, University of California Irvine School of Medicine
Joseph Kunisch, PhD, RN-BC, CPHQ, Harris Health System
Paul Kurlansky, MD, Columbia University Center For Innovation and Outcomes Research
Zhenqiu Lin, PhD, Yale-New Haven Hospital
Jack Needleman, PhD, University of California Los Angeles School of Public Health
Sean O'Brien, PhD, Duke University Medical Center
Jennifer Perloff, PhD, Brandeis University
Patrick Romano, MD, MPH, FACP, FAAP, University of California Davis Health
Sam Simon, PhD, Mathematica Policy Research
Alex Sox-Harris, PhD, MS, Stanford University
Ronald Walters, MD, MBA, MHA, MS, University of Texas MD Anderson Cancer Center
Terri Warholak, PhD, RPH, University of Arizona College of Pharmacy

Eric Weinhandl, PhS, MS, Fresenius Medical Care North America
Susan White, PhD, RHIA, CHDA, the James Cancer Hospital and the Ohio State University Wexner Medical Center

NQF Staff:

Dana Gelb Safran, SCD, President and CEO
Tricia Elliott, MBA, CPHQ, FNAHQ, Senior Managing Director
Tamara Funk, MPH, Director, Quality Measurement
Jill Herndon, PhD, Consultant
Sharon Hibay, DNP, BS, RN, Senior Consultant
Hannah Ingber, MPH, Senior Analyst
Gabby Kyle-lion, MPH, Coordinator
Elisa Munthali, MPH, Senior Consultant
Matthew Pickering, PharmD, Senior Director, Quality Measurement
Leeann White, MS, RN, Director, Quality Measurement

Also Present:

Mica Bowen, Research Assistant II, Brigham and Women's Hospital
Aileen Davis, PhD, Senior Scientist, Division of Health Care and Outcomes Research, Toronto Western Research Institute
Patricia Dykes, PhD, MA, RN, Associate Professor Of Medicine, Harvard Medical School; Program D Director of Research, Center for Patient Safety, Research, and Practice, Brigham and Women's Hospital
Stuart Lipsitz, ScD, Director, Biostatistics, Center for Surgery and Public Health, Brigham And Women's Hospital
Chang Liu, MPP, Project Lead/Data Scientist, Acumen LLC
Sriniketh Nagavarapu, PhD, Senior Research Director, Acumen LLC
Ronen Rozenblum, PhD, MPH, Associate Physician, Brigham and Women's Hospital; Assistant Professor of Medicine, Harvard Medical School; Director of Business Development, Center for Patient Safety Research and Practice; Director, Unit for Innovative Healthcare Practice and Technology
Stephanie Singleton, MPH, Project Coordinator, Brigham and Women's Hospital
Lisa Gale Suter, MD, Director, Quality Measurement Programs, Yale Core

Contents

## Proceedings

(11:02 a.m.)

## Welcome and Introduction

Ms. Elliott: Excellent, good morning, everyone. I wanted to welcome everybody to the Scientific Methods Panel fall 2021 Measure Evaluation Meeting. This is Tricia Elliott, I'm the Senior Managing Director here at NQF with responsibility for the support of the SMP Committee.

On our first slide here, some housekeeping reminders for day one. And I also wanted to let you know that the meeting is being recorded.

So this is a WebEx meeting. We're using the WebEx -- you can see some of the details in terms of the meeting length, meeting number, and password available. But if you're already hearing us, then you've made it this far. If there are challenges with audio, you can use the optional dial-in option also shared on this screen.

Please place yourself on mute when you are not speaking. We encourage you to use the following features. We have a chat box. You can message NQF staff or the group, the SMP group in its entirety. And also please use the raise hand to be called upon to speak.

There's two ways you can find the raised hand. At the bottom of the screen there should be a little icon image of a smiley face. If you click on that, there at the top of the call-all box, it says raise hand. Or if you happen to have the participant list open, you can hover over a name and raise a hand that way as well. There's a little hand icon that you can click on.

We will conduct the Scientific Methods Panel roll call in just a minute once we get through some introductions. And if you're experiencing any

technical issues, please contact the NQF project team at methodspanel@qualityforum.org, and they'll be able to help you troubleshoot via email as well. Or also let us know in the chat. Next slide, please.

Some -- oops, if you want to go back, Gabby. Thank you. Just a couple more housekeeping reminders. We do have a couple meeting breaks. Our first break is about 25 minutes, just so a heads up on that one if you want to grab some lunch or a quick snack before we enter into the afternoon portion of the meeting.

Voting is done by quorum. I mentioned the chat feature and raising hands. And muting and unmuting, there should be an icon at the bottom of your screen where you can mute and unmute. And if possible, we find that it's better to not use a speakerphone, either using a headset or microphone that's closer to your voice so that we can hear everybody well on the call.

And if you can introduce yourself during the discussions, particularly as you're entering into a large discussion. We are transcribing the discussion, so we want to capture everybody's comments.

And then we mentioned technical support. Please reach out via chat or the methodspanel@NQF.org or qualityforum.org for any issues you may be having.

And a question came through chat, do you want our videos most on or off. Whatever you're comfortable with. We prefer them on so we can see everybody, but a lot of times that is a drain with internet. So if you're having technical issues, feel free to, you know, have the video off.

So, some welcome, introductions, and disclosures of interest. So we'd like to start things off by introducing our CEO, Dana Safran. She's going to offer some introductory remarks for our committee. Dana.

Dr. Safran: Yeah, thanks so much, Tricia, and good morning, everyone. It's really my pleasure to welcome you to day one of the fall meeting of the Scientific Methods Panel.

I think I'll take just a moment. I know so many of you, but not all of you. So just take a quick moment to introduce myself and to share with you my background. So I am formally trained in the field of public health and quantitative methods.

I've been working in the field of quality measurement and improvement for about 30 years. Roughly the first half of that I was a measure developer, and roughly the second half of that I've been in a series of executive roles, more as a consumer of measures, using measures to drive improvements in quality outcomes, experience, and affordability.

And this moment in history of course is a tremendously important one for measurement in our country. And that really is a big part of why I took an interest in the NQF role and ultimately decided to come on board when offered.

There is, in addition to the critically important work that NQF has done for over two decades now around endorsement and maintenance, there is such important work that's motivated by our national attention to the urgency of addressing health equity.

Health equity has, you know, been in our strategic plans and mission statements for decades, and embarrassingly we collectively have made very little progress. This is an important moment that I think all of us recognize that conversation has become much more real. And NQF looks forward to a role that we can play in that space.

As well as in supporting the continued momentum public and private sector payers want to make toward payment reform. And the critically important

need for next generation measures and next generation data infrastructure to support measures and reduce burden while opening up new channels for the use of the subclinical data, leveraging, bulk buyer, and so forth.

So there's a lot of exciting work ahead. I couldn't be more thrilled, actually, to have this opportunity to lead this organization at this moment in time. And really want to acknowledge and appreciate the critically important work that this Scientific Methods Panel does.

Most of you probably recognize that SMP was first created in 2017 after some redesign work had been done using Kaizen method. And it was recognized that it would be critically important for a body such as this one to weigh in on the technical psychometrics, statistical aspects of measures in order to inform the Committee's work.

And so from then until now, this committee has played that extremely important role. And in fact we've seen that over time the percentage of measures that are needing to come before this panel because of their technical elements has been increasing. So today's discussion of course will be a discussion around evaluating measures. Tomorrow the SMP will turn to its advisory role, and specifically dealing with some issues around measure reliability and some other aspects that I know are on your agenda.

So I just really want to take this moment to thank each and every one of you. I know this is not only a big chunk of your week when you have a two-day meeting, but the preparations for this and the ongoing work really are significant, and we appreciate that.

And in particular want to voice my appreciation for our Co-Chairs, Dr. Nerenz and Dr. Teigland. So thank you much to the two of you and this whole panel. Let me turn it back to you, Tricia.

Ms. Elliott: Thank you so much, Dana. We really appreciate your comments and kind of getting us started on our two-day journey here. So we're excited to undertake the work with our esteemed colleagues here. Next slide, please.

By way of quick introductions, I just want to share with you the folks from NQF that are staffing this call today and prepared all of the information that you've been reviewing in preparations for the meeting.

As mentioned, my name is Tricia Elliott, I am the Senior Managing Director here at NQF.

We also have Mike DiVecchia, who's our Senior Project Manager; Hannah Ingber, our Senior Analyst; Gabby Kyle-Lion, our Coordinator; Sharon Hibay is a consultant on the project, as well as Elisa Munthali is also a consultant, and Jill Herndon, a consultant's been supporting the project as well. So, many thanks to the NQF team for preparing everything today and we're very excited for a great meeting today. Next slide, please.

At this point, we're going to pause to do introductions, or it's going to be combined with disclosure of interest. So I want to thank everyone for their time today, and we'll combining, as I mentioned, the introductions with the disclosure of interest.

So you've 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 and meeting today. 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 the Committee's work.

Just a few reminders. You sit on this group as an individual. You do not represent the interests of your employer or anyone who may have nominated you for this committee. You are interested -- 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 disclose does not mean that you have a conflict of interest. We do verbal disclosures in the spirit of openness and transparency.

Now we'll go around our virtual table stating -- starting with our Committee Co-Chairs. I'll call your name. When I do so, please state your name, what organization you're with, and if you have anything to disclose. If you do not have disclosures, please just state that I -- or make the statement I have nothing to disclose.

To keep us moving along, if you experience trouble unmuting yourself, please raise your hand so that staff can assist.

So I'll start with our Committee Co-Chair, David Nerenz.

Co-Chair Nerenz: Yeah, good morning. David Nerenz, Henry Ford Health System. Nothing to disclose today.

Ms. Elliott: Excellent, thank you, David. And Christie Teigland will be joining us later, so we'll circle back with her. Matt Austin.

Member Austin: Yeah, good morning to everyone. Matt Austin from the Johns Hopkins Armstrong Institute for Patient Safety and Quality, and I don't have anything to disclose.

Ms. Elliott: Great, good morning, Matt, thank you.

Bijan Borah. We'll circle back. John Bott.

Member Bott: Hi, John Bott, independent contractor working with the Alliance and The Leapfrog Group, and I have nothing to disclose. Thank.

Ms. Elliott: Excellent. Thank you, John. Daniel Deutscher.

Member Deutscher: Hello, this is Daniel. I'm a research scientist with Net Health Systems in the US, and the Maccabi Healthcare System over here in Israel, and I have nothing to disclose for today.

Ms. Elliott: Thank you so much. Lacy Fabian. Marybeth Farquhar.

Member Farquhar: Good morning. My name is Marybeth Farquhar, I am the Executive Vice President for Research Quality and Scientific Affairs of the American Neurological Association, and I have nothing to disclose.

Ms. Elliott: Excellent, good morning, thank you. Jeffrey Geppert.

Member Geppert: Hello, Jeffrey Geppert, Battelle Memorial Institute, and nothing to disclose for today.

Ms. Elliott: Thank you. Larry Glance.

Member Glance: Good morning, everybody, I'm Larry Glance. I am at the University of Rochester, and I have nothing to disclose for today.

Ms. Elliott: Thank you. Joseph Hyder. Sherrie Kaplan.

Member Kaplan: Sherrie Kaplan, University of California Irvine School of Medicine. I recently received an additional grant from the Patient-Centered Outcomes Research Institute, where we're now implementing child health rating inventories to study the impact of improved -- on improved diabetes care of new technologies.

A supplement we just got was to use those same measures in both children and now adults with diabetes to study the impact of telehealth. And that's all I have to disclose.

Ms. Elliott: Great, thank you, Sherrie. Joe Kunisch.

Member Kunisch: Good morning, Joe Kunisch. I'm the Vice President of Quality Programs at Harris Health System in Houston, TX, and I have no disclosures.

Ms. Elliott: Great, thanks, Joe. Paul Kurlansky.

Member Kurlansky: Yeah, hi, Associate Professor of Surgery at Columbia University Center for Innovation Outcomes Research in the Department of Surgery, and I have nothing to disclose.

Ms. Elliott: Thank you, good morning. Zhenqiu Lin.

Member Lin: Yeah, hi. I'm a Senior Research Scientist at Yale University, and I do work under contract with CMS to develop a quality measure.

Ms. Elliott: Thank you so much, good morning. Jack Needleman.

Member Needleman: Good morning. Jack Needleman, the UCLA School of Public Health, and nothing to disclose.

Ms. Elliott: Great, thank you so much. Eugene Nuccio. I believe he's unable to attend today. Sean O'Brien.

Member O'Brien: Hi, I'm in the Biostatistics Department at Duke University Medical Center, and I have no disclosures.

Ms. Elliott: Thank you. Jennifer Perloff.

Member Perloff: Hi, Jenn Perloff. I'm a Brandeis University, and I have no disclosures.

Ms. Elliott: Great, thank you. Patrick Romano.

Member Romano: Patrick Romano, I'm at UC Davis Health in Sacramento, CA. And like Dr. Lin, I work extensively as a measure developer under subcontracts with CMS and AHRQ, mostly related to hospital measures of hospital harms or patient safety.

I also was involved as a consultant with three of the measures that were submitted for this cycle to SMP but were not pulled for review at this meeting.

Ms. Elliott: Great, thank you, Patrick. Sam Simon.

Member Simon: Good morning, everyone. Sam Simon, I'm a Senior Director at Mathematica. Mathematica holds measure development contracts with CMS.

Ms. Elliott: Okay, thanks, Sam. Alex Sox-Harris.

Member Sox-Harris: Good morning everyone. Sox-Harris, I'm a professor in the Department of Surgery at Stanford University and a Healthcare Researcher at the VA. I have no disclosures.

Ms. Elliott: Thank you. Ron Walters.

Member Walters: Hi, Ron Walters, I'm a medical oncologist in Anderson Cancer Center. I have nothing to disclose from either perspective, except that I appear to be on the group that pulled six measures, as opposed to the other group that pulled one. There must be a bias there somewhere.

Ms. Elliott: Interesting observation, thank you. Good morning. Terri Warholak. Moving along, Eric Weinhandl.

Member Weinhandl: Yeah, good morning. Hi, Eric Weinhandl, Senior Epidemiologist at Hennepin Healthcare. I do have consulting relationships with Fresenius Medical Care as well as a few dialysis device manufacturers. So to the extent the measures arise in that space, I don't think they do at this meeting, I occasionally have some conflicts.

Ms. Elliott: Okay, thank you. Susan White.

Member White: Hi, good morning. Susan White, I'm with the Ohio State Wexner Medical Center. I don't have any conflicts with the measures that we're looking over today. Thank you.

Ms. Elliott: Great. I'm going to circle back on a couple folks we didn't hear from to see if they've joined the meeting. Bijan Borah. Lacy Fabian. Joseph Hyder. One second. And then I got a note Terri Warholak will be joining, logging in around noon.

Okay, we have 20 out of 26 members present, so we are going to move forward. Could I have the next slide, please.

Meeting Overview

Okay, the next portion of the agenda I'll be getting into the meeting overview. So we'll review the agenda real quick for day one. So we're working our way through the welcome and introductions, and we completed the disclosure of interest.

We'll be providing some evaluation updates from the spring 2021 and fall '21 cycles. We'll do a process overview and evaluation reminders. We'll have a quick break from 12 to 12:30, and then we'll get into the fall '21 measure evaluations, allow an opportunity for public comment, and then we'll be adjourning.

So the meeting materials provided for you today include the discussion guide. So this is a synopsis document of the scientific acceptability content. So this the reliability and validity requirements for all the complex measures in the measure cycle that's being evaluated by this committee.

Each measure includes pertinent information from the submission, the SMP reviewer feedback, related developer responses, and identification of measures

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

Appendix B within the discussion guide is additional information provided by measure developers. In addition we have provided background materials, including testing -- the 2011 testing task force report, the 2021 NQF measure evaluation criteria and guidance, as well as the SMP measure evaluation guidance. Next slide please.

At this point I'm going to hand things over to Hannah Ingber to take us through the next couple of slides.

Ms. Ingber: Thanks, Tricia. Morning, everyone. I'm just going to go over last cycle's stats a little bit. In the past, SMP members have requested that we sort of give an update on last cycles' measures, so we're going to do that here.

As you can see, 29 measures were evaluated by the SMP last cycle, and 13 of them, about half, were discussed at the meeting. The final results from the SMP meeting last cycle was that 23 of the 29 measures passed and were evaluated by the standing committees.

There were two where consensus was not reached, and there were two that did not pass, and there were also two that were withdrawn mid-cycle. So the standing committees revoted on the scientific acceptability for two of the 29 measures, and on the next slide we'll show you a little bit more about those.

So you can see here that we -- consensus was not reached for validity on these two measures, 3621 and 3622, so that was why the standing committee revoted on them. And they were both recommended

for endorsement after the standing committee's measure evaluation meeting. Next slide, please.

And I'll also note that some post-comment meetings are still happening, so these statistics are up to date as of now, but may get one or two measure changes after the Consensus Standards Approval Committee meets in November and December.

So again, in spring 2021, we had 29 measures. Of those, 23 passed, two of them did not pass, and two of them were consensus not reached, and two of them were withdrawn.

So seven of those 29 ended up not moving on to the standing committees. They were withdrawn from the cycle, so that means our denominator on the bottom right box, 19 over 22, is 22. That's how we get that denominator. So 19 of them were in agreement with this SMP. And so for the three -- or sorry, for the two where the -- okay, so I mentioned that some were withdrawn, that was part of the reason for the denominator change. But for three of the 22, no vote was taken for scientific acceptability because the measure didn't pass on evidence. So it's not a matter of much -- as much of disagreement as the evidence was presented and the standing committee did not pass the measure on that criteria.

So that's our update for the spring 2021 measures. If we could go to the next slide, I will pass it back to Tricia unless there are any questions from our SMP members.

Member Austin: So real quick, I have quick question. If you -- on that previous slide it sounded like really all 19 measures that did get voted on by the standing committees agreed with the SMP recommendation. Do we know if for prior cycles if that is a similar sort of issue with that, you know, some of the dropoff is actually that the measure wasn't actually voted on by the standing committee?

Ms. Ingber: Right, yeah. You can see the previous cycle statistics on this bottom row. But it is a bit of a mix of whether it was because of the standing committee not getting to scientific acceptability or disagreeing with the methods panel. So we can provide those more detailed statistics to you if you'd like, but for the most part it is a matter of not getting to that criteria.

Member Austin: Yeah, my thought is maybe just for future updates, it doesn't obviously need to be something you go back and do retrospectively. But at least my understanding was I think to the SMP, it was partly trying to understand where the standing committees came to a different conclusion than the SMP. So I think if we could limit it to just those that they actually voted on, that might give us a clear picture of where there's agreement and disagreement. Thank you.

Ms. Ingber: Sure, thank you.

Ms. Elliott: I was on mute. Before I move into the fall 2021 cycle, I believe Lacy Fabian was able to join us. Lacy, if you could state your name, your organization, and if you have anything to disclose. If you're speaking, Lacy, I believe you're on mute. We can't hear you.

Okay, we'll circle back. Seems to be an audio issue there, but Lacy is on the line I think. Okay, next slide, please.

## Evaluation Updates

So for the fall 2021 evaluation cycle statistics, we had 12 complex measures that were assigned to the SMP. Eleven were new measures, so and one was a maintenance measure. Two subgroups of the -- of 12 to 13 SMP measures were each assigned six measures.

Eight measures passed reliability and validity. Two measures were consensus not reached on reliability

or validity. One measure did not pass on reliability, two measures did not pass on validity. Three measures were withdrawn prior to SMP preliminary review. And seven are slated for discussion.

On the righthand side is a summary of the types of measures, so four outcome, three intermediate clinical outcome, four PRO-PMs, and one process measure.

I believe there's a hand raised. I'll pause here for a second. Sherrie Kaplan, did you have a question?

Member Kaplan: Yeah, it was just a curiosity, because I'm wondering if there are some criteria that are being considered by NQF on when to bring a new measure before NQF. And you know, some criteria for what it takes. That I notice that with all these new measures, some of them are in different stages of maturity and development.

So are you all considering some criteria for when to bring a measure forward and what those standards might look like?

Ms. Elliott: That's a great question. So at this point in time we have all of our criteria within the MIDS tools, which is the new portal which the new portal that is used by the measure developers to submit the measure. So we do a complete list check and then start the measure through the process. And if the measure developer completes all of that content that's requested, then the measure can go through the cycle.

As it passes through, there may be points in time that a measure developer makes the decision that they want to, you know, pull back and move it to another cycle if it's not quite ready. So those discussions are a little bit ongoing.

I believe the final submission deadline is early November, so there could be some changes then based on the process. The measure developer could

make that decision. But it's a great question that we can pursue further to see if there's additional criteria or guidance that we can provide to help with the process.

Member Kaplan: Thanks.

Ms. Elliott: So if I can summarize your question, it's almost like readiness of the measure type of thing? Okay.

Member Needleman: This is Jack, if I can just follow up on that, because it's going to be relevant to a couple of the measures that we discuss today. Measures go through development cycles, and very early alpha testing. And then potentially go through a much larger scale beta testing with a lot more groups, a lot more data.

A number of the measures that we've got today look like they're still at that alpha stage, a very select, very narrow group of providers that were analyzed with a lot of limitations on the analysis that can be done around reliability.

And the question is -- one of the questions is whether endorsement, that that stage so that that developer has imprimatur to go out and recruit a larger body to continue developing the measure is appropriate or whether we want to tell the developers, no, you got to do that stuff before you get endorsed.

Ms. Elliott: So I think Elisa, do you have any history on this? Has this discussion -- newer to the process, so I just wanted to reach out to Elisa, because I know there might be a little history on that in terms of prior discussions.

Ms. Munthali: Yeah, it's a great question. It's one that NQF has struggled with for some time. But over the years, what we did try to do, and I think Tricia mentioned it as part of the evaluation criteria is to provide very detailed guidance. So developers and

committees have a roadmap on how, you know, which measures they should bring forward and how they should be evaluated.

We did at one point also create a what's -- what good looks like document for various measures that came forward to NQF. And I think that's where perhaps, you know, we can look a little further in to see what it -- what, you know, revisions we may need to make, given where we have moved in measure development, even in just a few years.

So I think something like that that gives developers something tangible about, you know, what ideally NQF is looking for to accompany the evaluation and the guidance that we already have.

Ms. Elliott: Thanks, Elisa. I think those tools are helpful and will help us further the discussion.

I'm going to pause here for one more second just to make sure there's no more hands raised. And I believe Christie Teigland was able to join us. Christie, are you able to unmute?

Ms. Teigland: Yes.

Co-Chair Teigland: Yes, hello.

Ms. Elliott: We can hear you, good morning.

Co-Chair Teigland: I am here. Happy to join, sorry I'm a few minutes late. I'm catching up. I did my first conference, AMCP in Denver last week, and so I'm getting back and catching up today. But happy to be able to join the rest of this meeting.

Ms. Elliott: Excellent, and just for the formality, do you have anything to disclose?

Co-Chair Teigland: I do not.

Ms. Elliott: Excellent. Thank you, Christie. And we're going to try and see, Lacy Fabian, are you able to unmute? Okay, we'll circle back and see if we can

help. Oh, Lacy, are you unmuted?

Member Fabian: Hi, yeah, I didn't want to interrupt the meeting a third time just to double mute. Lacy Fabian, with MITRE, nothing to disclose.

Ms. Elliott: Excellent. Thank you so much, Lacy, I appreciate your persistence in getting unmuted there. Excellent. Okay, next slide, please.

So for fall '21, the measures slated for discussion are outlined here on this slide. We have a number of measures under the surgery group, one under primary care and chronic illness, and from Subgroup 2, one measure under patient safety.

Process Overview and Evaluation Reminders

So the next section of the agenda is to go through some of the process overview and reminders. So next slide, please.

So the overall ratings for the measures are high, moderate, low, and insufficient. So a score of high is testing is -- the score-level testing is required. A measure may be eligible for a high, but the sampling method results may warrant a moderate rating.

Moderate, the highest eligible rating if only data element testing or face validity -- excuse me, face validity testing is conducted. A measure may be eligible for moderate, but the sampling method results may warrant a low rating.

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

And insufficient is used when the reviewer does not have sufficient information to assign a high, moderate, or low rating. Examples, unclear specifications, unclear testing methodology, or not conducting criteria required in the testing.

So the task force - no, I think we're good. Next slide. Sorry, Gabby, through you off cadence there.

So this next slide speaks to the quorum. A meeting quorum is met with 66% of active SMP members in attendance, which we have achieved today. Achieving consensus is calculated from the percent of quorum members during a vote. SMP scientific acceptability, for example the reliability and validity criteria evaluation results.

So pass recommended is greater than 60% yes of quorum votes. So that's the high plus moderate ratings. Consensus not reached, CNR, is 40-60% yes of quorum votes, inclusive of 40% and 60%. Does not pass or not recommended is less than 40% yes of quorum votes.

Differences in testing requirements by measure type. Health outcomes, intermediate clinical outcomes, cost and resource use, and structure and process measures. For both reliability and validity, NAF requires either patient or encounter level, previously known as data element level, testing, or accountability entity level, previously known as measure score level, testing for these new measures.

Both testing types are preferred, yet not currently required. Impacts to rating as previously described. An exception here is face validity testing of the computed measure score for a new measure is accepted at the accountable entity level.

If patient encounter level validity testing is provided, we do not require additional reliability testing. In this case, use the rating you'd give for validity as the rating for reliability.

Submissions that accept patient encounter level validity testing for patient encounter level reliability testing is occurring is less frequently in recent measure cycles. Thanks, Gabby.

Differences in testing requirements by measure type. So this slide addresses the instrument-based measures, including the PRO-PMs. For reliability and validity testing, testing is required at both patient encounter, the data element, or accountability entity measure score levels for initial endorsement evaluation.

Patient encounter level testing must be conducted for reliability and validity of the multi-scale -- multi-item scales at the patient level. Accountable entity-level testing must be conducted for reliability and validity testing of the actual performance measure at the level of analysis as defined in the measure specifications.

Face validity testing of the computed measure score is accepted at the initial endorsement evaluation in lieu of empirical, accountable entity-level validity testing.

Differences in testing requirements continue for composite measure. So NQF provides specific guidance and definitions for the composite measures. Components of the composite measures should have their own properties of reliability and validity.

NQF does not consider multi-item scales in survey -- surveys or questionnaires as composites. NQF does not consider multiple component measures without a single performance rate and multiple component performance rates as composites.

Accountable entity-level reliability testing of the composite is required. Demonstrating reliability of individual components alone is not sufficient to pass the criterion. Accountable entity-level validity testing is not required until maintenance.

Additional specific acceptability criterion is used for composite measures. Empirical analyses supporting the composite construction, including the value of their components to the composite and the

component aggregation and weighting consistency to composite quality construct.

Some testing and evaluation reminders. All testing must align with the specifications. This is not a new requirement that NQF is more rigorously upholding this requirement, particularly for the level of analysis testing and minimum sample sizes. If multiple levels of analysis are specified, each must be tested separately.

NQF's requirements permit passing some or all levels of analysis for a measure. Occasionally there are several performance measures included under one NQF number. Each measure must be evaluated separately. Some measures may pass and others may not pass.

I'm going to pause here for a second, I think there -- I believe there's a hand raise. Sherrie, did you have a question before I move on?

Member Kaplan: Well, two questions. One is the line number -- if multiple levels of analysis are specified, each must be tested separately. And then the next line says NQF's requirements permit passing some or all. And the must and some or all are confusing on that line.

And then I'm back to the sort of measurement model, back to the composite measures for a minute. We talked at one time about formative versus reflective models of measurement and the changes those make in the assumptions you make about how the elements of a component are related to each other, which in turn changes the appropriate methods.

So I'm wondering if somewhere in the parking lot we can, you know, resurrect that discussion.

Ms. Elliott: Right, thank you. We can definitely capture that piece. I'm going to call upon the NQF team. Do we have any clarification we can offer

Sherrie on the -- the alignment with the specifications?

Ms. Ingber: I'm happy to jump in here unless other --

Ms. Elliott: Go right ahead, Hannah, thank you.

Ms. Ingber: Yeah, so the testing must be for each level analysis. But it is possible to pass part of a measure if the testing is acceptable for one level of analysis and not another. Does that make more sense?

Member Kaplan: So is one the testing is okay but the results are not okay, is that the distinction I'm hearing? Because otherwise I'm still confused.

Ms. Ingber: If I understand you correctly -- yeah, go ahead.

Co-Chair Nerenz: Hi, yeah, let me try to paraphrase it and see if I can get it. And I'm just sort of reflecting the words. Let's say you've got a measure where the developer says this could be used at the individual physician level, physician group level, or ACO level, so there are three levels.

I think the rule says you must test at all three levels. The rules do not say, however, that you must pass at all three levels. That's basically I think thrown to us at the Scientific Methods Panel. If reliability looks good, for example, at two out of the three, then it's our call what rating to give it.

So and I'm just reflecting the words on the page, that -- and I'm seeing Hannah nod. Is that a fair rate statement?

Ms. Ingber: Yes, thank you so much.

Co-Chair Teigland: Can I just add though that I've run into a few cases like this and I'm trying to get my video on. I've ran into a few cases like this where they do test or it at various levels, and they

show that it's, yeah, you know, it's reliable at, you know, the provide group level but not at the physician level, unless you have a minimum denominator.

But we can't -- we have to vote the whole thing because it's submitted as one measure. So I will sometimes write in there, you know, yes, acceptable, assuming they follow these minimum threshold guidelines for the denominator if it's applied to a provider. But you don't know if that happens, right.

So it is -- I don't think the forms are set up for us to approve or disapprove it at those different levels. Perhaps we need to, you know, make some of those distinctions when we do find differences in reliability for different applications of, you know, provider group versus provider.

Ms. Elliott: Great, thank you for that, Christie. Any other comments or questions before we move on? I think we have another committee member who was able to join. Let me see if audio is up. Terri Warholak appears to have joined. Could you share your name, organization, and if you have anything to disclose.

Member Warholak: Sure. Terri Warholak, University of Arizona College of Pharmacy, and I have nothing to disclose.

Ms. Elliott: Excellent, thanks for joining, Terri. Next slide, please.

So we have some additional reminders to share. Consideration for risk adjustment is required for all outcome resource use and some process measures. Inclusion or exclusion of certain factors in the risk adjustment approach should not be a reason for not passing a measure.

Concerns with discrimination, calibration, or overall method of adjustment are grounds for not passing a

measure.

In the absence of risk adjustment for outcome, resource use, and some process measures, a strong rationale or data for excluding must be provided. For all measures, incomplete or ambiguous specifications are grounds for not passing a measure. But remember that there is an option to get clarifications, although this must be done early on.

Empirical validity testing is required at the time of maintenance evaluation. If not possible, strong justification is required and must be accepted by the standing committee.

Additional reminders. The SMP articulated additional guidance for submissions. Provide greater detail when describing testing methodologies and results. Provide more than one overall statistic when conducting signal-to-noise reliability testing.

Provide greater detail in description of construct validation describing hypothesized relationships. Why examining hypothesized relationships would validate the measure. Expected direction and strength of the association. Specific statistical test used results, results interpretation, how the results related to the hypothesis, and whether the results assist to validate the measure.

Lack of two of number two on this list and number three should not be grounds for not passing a measure.

All measures reviewed by the SMP can be discussed by the standing committees. Standing committees will evaluate and make recommendations for endorsement for measures that pass SMP review, measures where the SMP did not reach consensus. And measures that do not pass the SMP may be pulled by the Standing Committee member -- by a Standing Committee member for further discussion and revote if it is an eligible measure.

It appears, Patrick, did you have a question?

Member Romano: I did. Could you go back, I think it's two slides, to the risk adjustment. I just wanted to clarify a point here. So the first bullet point about inclusion or exclusion of certain factors in a risk adjustment approach should not be a reason for not passing a measure.

First, I think if I'm correct that this is specific to the SMP discussion. And I think it's motivated by a concern that the SMP is not the repository of clinical wisdom regarding what should be considered in the risk adjustment or what should not be.

However, the question I think comes up for some of the measures we're discussing today, in terms of the general approach, and specifically whether certain types of variables should be considered as risk adjusters or not. And I'm wondering if I could hear some further explanation of that question.

In other words, the choice of what types of factors to include in the risk adjustment obviously is tied to the issues of discrimination and calibration, the overall method, so forth.

So I think the idea is that we're not supposed to get into the details of this condition was used for risk adjustment and that one wasn't. But perhaps some clarification could be provided about the motivation for that first bullet.

Ms. Elliott: Sure. Sharon, did you want to?

Dr. Hibay: Thank you, Tricia. Patrick, this is Sharon Hibay. Are you asking about also the types of variables? So clinical, demographic, social risk, functional. We know that historically we have been utilizing both the demographic and clinical or health status, and we are working on the social risk, and also the functional.

We have a project with risk adjustment, etc. So I

think that you're asking about those specifically. Again, one of the measures talks about a very specific variable that's a clinical one. I don't think you're trying to get into that minutiae, but you're asking for a little bit more guidance on the more bigger bubbles of the types of variables to be included in the risk model. Is that what you're saying?

Member Romano: I'm trying to clarifying if we're precluded from discussing anything about what variables or what types of variables are in the model, and if so why, since that's kind of integral to the validity issue.

Dr. Hibay: Yeah, so, and Elisa and obviously Tricia, if you could also assist with this. So you know, longstanding, again, demographic and clinical have been part of risk adjustment models forever. I'm not sure where we are. I don't think we have totally finalized that all models have to also consider for social risk and functional risk at this time.

I also wonder if Matt Pickering is, Dr. Pickering is on the call as well, to see if he could provide any feedback. I don't see him on the call.

Ms. Elliott: Yeah, no, we're still working through some of that in terms of the implementation and impact of some of the risk adjustment models. But to your point that the bullet that you called out, Patrick, is specific to the SMP and, you know, the guidance that we've used in the past is when the SMP has trouble with a particular variable in a risk adjustment model or use, they should probably pass it to the Standing Committee, but they're not -- we don't -- we're not preventing any discussion on this, it's just not a voting issue for the SMP.

Does that help to clarify a little bit?

Member Romano: Yes, yeah.

Dr. Hibay: Okay.

Member Romano: I mean, as long as there's a clear mechanism by which our concerns about a variable would be transmitted, I guess, to the Standing Committee. I mean, so just to be specific, I think we'll be discussing some risk adjustment models that include race as a specific factor. And that's a highly controversial issue, the inclusion of race as a risk adjustment variable, especially in process -- process measures.

So that's, you know, so I might interpret that as something that, you know, absolutely should not be done. Somebody else might interpret it as something that's outside our scope to discuss. And so that's -- I'm just trying to clarify.

Ms. Elliott: Yeah, so the discussion would get captured as such. Right.

Awesome, so I think we left off -- we did additional reminders. I think we move -- oh, Paul Kurlansky, did you have a question or comment?

Member Kurlansky: Yeah, no, I want to thank Patrick for pointing this out, because I probably have less of an issue with the second -- with the first bullet than I do with the second bullet. In other words, if we find that a model is completely, you know, poorly constructed and really does not adequately convey risk adjustment, then the entire basis for the validity of the metric may be in question.

But here it's saying that if the discrimination calibration which are the measures of the adequacy of a risk adjustment model, cannot be your grounds for not passing the measure. It would seem to me that would be a very solid grounds for not passing the measure.

Co-Chair Nerenz: I think it says they are grounds, second bullet.

Member Romano: That (Simultaneous speaking.) in

a very peculiar place, but yes, I'm with David on this one. It says that we can reject a measure that doesn't meet our calibration discrimination or adjustment standards.

Member Needleman: But it, my point is it's sort of tied also with the first bullet, because usually the reason it doesn't pass is because it's omitting certain types of risk factors. But that's fine, we can focus on the second bullet.

Co-Chair Nerenz: Sure. That's right, Paul. But I -- at least my read of it says that badly calibrated or badly constructed risk adjustment is or are grounds for rejection. It's all pretty -- you got to read it carefully.

Member Needleman: Got it, okay, thank you.

Member Sox-Harris: This is Alex. Just to jump in on this point just to, one issue we've run across in the past is not the exclusion of variables in a risk model but the inclusion of variables that don't belong in there, for example, that happen after the, you know, the index event.

So, it's, that would be a case I think where if there are variables that happen after the, you know, the index event, then that speaks to the validity of the risk adjuster and we should be able to consider that.

Ms. Elliott: There's a -- or a comment in the chat from Zhenqiu. Zhenqiu, did you want to share your comment, please?

Member Lin: Yeah, I think this could follow on what Patrick just mentioned, right. So if someone used some variable that capture quality of care for risk adjustment, then shouldn't it be ground for not passing? Like if you use a in-hospital complication to risk adjust for 30-day mortality, obviously you actually make the prediction better. But that's something you should avoid.

Co-Chair Nerenz: Now that, just my own thoughts here. You know, we all like to work with really bright line distinctions and say on this side of the line we can act and decide and on the other side we don't. But this seems one of those that's kind of a gray area where, on the one hand, yes, we're talking about a variable or variables that are in or out of the model. Presumably that falls to the Standing Committee.

But all of us are looking at it, and Zhenqiu, just in the example, would say, you know, just, this isn't about clinical wisdom, this is just about general board rules about how one constructs risk adjustment models.

And if that appears, we have to say we're not confident in the results of the risk adjustment modeling, and therefore we have concerns about validity and reliability. And it's really hard for us to say we're just going to ignore that and pass it on to the Standing Committee.

We just may always live with a little bit of ambiguity of what's at the border of what's clinical wisdom on the one hand belongs to the Standing Committee, and what's about the structure and the general rules of model-building that belong to us. So it's, it may never be a complete, absolute bright line distinction.

Ms. Elliott: Appreciate that clarification and all the comments. I think we have a couple more slides to get to in this section. So Slide 28, thanks, Gabby.

So committee consideration of measures that do not pass the SMP. So the eligibility will be -- will be determined by NQF and SMP co-chairs.

So measures that did not pass the SMP due to the following will not be eligible for a revote: 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 SMP to apply the criteria, and appropriate levels of testing not provided or otherwise did not meet NQF's minimum evaluation requirements. Next slide, please.

Okay, at this point, we are transitioning into review of our first measure. I want to note we have 35 minutes is given to discussion of each measure. So the discussion here, the discussion process includes the measure will be discussed by the SMP and our determined during the SMP measure review activities. Staff will -- NQF staff will briefly introduce the measure. SMP, there is an SMP member who has been designated as the lead discussant and will summarize key concerns.

Other SMP subgroup members are invited to comment. Developers are given two to three minutes for an initial response and may respond to SMP questions. Discussion are open to the SMP and proceed by individual criterion.

Recused members cannot discuss measures where conflicts are identified.

The voting process. Voting is conducted synchronously, virtually, and confidentially via Poll Everywhere. Voting occurs following each criterion discussion. SMP subgroup members only vote on measures they were assigned. Recused SMP members cannot vote for the measure where conflicts are identified.

Subgroup voting results taken during the meeting are the official SMP vote. Measures that are not pulled for discussion will pass in a consent calendar vote. Next slide.

So at this point we are going to do a voting test, and I'll hand things over to Hannah.

Ms. Ingber: Thank you, Tricia.

So, shortly in your inbox you should be receiving the voting link so that we can conduct the voting test. We're going to put up a test question for folks to respond to so we can make sure everything is working all right.

And I'm going to share that now.

(Pause.)

Ms. Ingber: So, we're asking you to let us know what your favorite Halloween candy is of these two. Do you prefer Kit Kat or Gummy Bears?

Co-Chair Nerenz: Hannah, do you want everybody doing this or just Subgroup 2?

Ms. Ingber: Oh, right. Yes, we want everyone to do this. And it's especially important for members of Subgroup 2, but we want to confirm that everyone is able to use the software.

Member Deutscher: Hannah, excuse me. I'm not sure where to find the link for the vote.

Ms. Ingber: You should have it in your email. But I can send it to you again.

Co-Chair Nerenz: Oh, in the email.

Ms. Ingber: Uh-huh.

Co-Chair Nerenz: Yes, I do see it now. Thank you.

Ms. Ingber: Great.

Member Needleman: I would appreciate having the link sent to me again.

Ms. Ingber: Sure. That was Jack? Yes.

Member Needleman: Yes.

Ms. Ingber: Can do.

Member Romano: And what happens? It says that

the presentation is underway. As soon as the activity is active, you'll see it on the screen.

Ms. Ingber: Ah. It should be active for you. Let me make sure.

Member Romano: Maybe it's only -- Ah, there we go. Okay.

Ms. Ingber: There we go. Apologies.

Co-Chair Nerenz: On mine there's a little bit of delay. There's kind of this welcome screen and then it just, it transitioned to the question. It took it a while.

Ms. Ingber: Okay. I think we're waiting on two more.

Member Needleman: Yeah. I haven't gotten the link yet. The UCLA mail service seems to occasionally not let perfectly reasonable email through, or delays ridiculously.

Oh, just got it.

Ms. Ingber: Oh, great.

Member White: And I -- Oops, never mind. I was going to say I have the same, waiting for it to begin. But I got it.

Ms. Ingber: All right. Let me just confirm.

I think we're waiting for just one more person. And you can feel free to chat me privately if you're, if you're having any trouble.

Member Romano: And I assume you can chat with us if you know that we're the one whose vote doesn't come in?

Ms. Singleton: Hi. This is Stephanie from Brigham. And I am not receiving the facts.

Ms. Ingber: Oh. This is just for SMP members.

Ms. Singleton: Perfect. Okay, thank you.

Ms. Ingber: Sure.

I will check to see if I can help anyone. Let's see.

(Pause.)

Ms. Ingber: I appreciate everyone's patience.

Ms. Elliott: It's good to work out these kinks before we get to other votes. So, thanks, Hannah, for double checking behind the scenes.

I did get a comment that candy corn is not on the list, so. If you'd like, add candy corn.

Ms. Ingber: I will say that we have reached forum for the voting. So, we can work on the side to identify who we can help.

Member Perloff: And the poll results show?

Ms. Ingber: Oh, I'm sorry. Yes, it would probably be good to see this.

Member Perloff: Yes.

Ms. Ingber: Let's see. Okay. The poll results show that the majority of the panel prefers Kit Kats to Gummy Bears. We have 19 votes for Kit Kats and 3 votes for Gummy Bears, for a total of 22 votes. Thank you, everyone.

Ms. Elliott: Awesome. Thank you, Hannah. Good to know when we're back in person we can stock up our candy dishes with Kit Kats and make the majority of us happy.

Wanted to, before we move into the next section, just wanted to double check to see if Bijan Borah or Joseph Hyder, if either of you are on the call. We have a few call-ins that don't identify names, so we just wanted to double check.

Okay. And, Gabby, if you could, or, Hannah, if you

can pass the presenting back to Gabby. Excellent.

0689 Percent of Residents Who Lose Too Much Weight (Long-Stay) (CMS)

So, at this point I'm going to hand things over to Tammy Funk, who is a director here at NQF. And she will be walking through Measure No. 0689, Percent of Residents Who Lose Too Much -- excuse me, Who Lose Too Much Weight.

Tammy, are you on the line?

Ms. Funk: I'm here, Trisha. Thank you for that.

Good morning, or I guess good afternoon to the Methods Panel. And thank you very much for reviewing this measure on behalf of the Patient Safety Team. And we look forward to your discussion around it today.

So, this measure, 0689, Percent of Residents Who Lose Too Much Weight, Long-Stay, is developed and stewarded by Acumen and the Centers for Medicare and Medicaid Services. This measure is undergoing maintenance review, and it is an outcome measure.

A brief description for you.

This measure reports the percentage of long-stay nursing home residents with a targeted minimum data set assessment that indicates a weight loss of 5 percent or more of the baseline weight in the last 20 days, or 10 percent or more of the baseline weight in the last 6 months which is not a result of a physician-prescribed weight loss regimen.

The baseline weight is the resident's weight closest to 30 or 180 days before the date of the target assessment. Long-stay residents are identified as residents who have had at least 101 cumulative days of nursing facility care.

This measure is assessed at the facility level, and it is not risk adjusted.

The data source for this measure is the minimum data set of 3.0, and the correction instrument is the resident assessment instrument.

So, for reliability, SMP reviewers passed this measure with a moderate rating. Reliability testing was conducted at the accountable entity level, which is the facility. And for critical data element reliability, the developer completed a kappa analysis of gold standard nurse to facility nurse.

For performance measure score reliability, the developer conducted both split-half reliability and usual beta binomial signal-to-noise reliability testing.

The split-half reliability testing correlation was positive, and the relationships was moderate.

For validity, this measure was consensus outreach.

As with reliability, testing was conducted at the facility level.

The developer conducted critical data element testing, relying on previous studies that have looked at inter-rater agreement. So, they examined a national validation of the minimum data set 3.0 that tested the criterion validity of the items by examining the agreement between gold standard nurse assessments and facility nurse assessments based on kappa statistics. They also tested the measure score using, first, the correlation with other measures of nursing facility quality, including facility CMS five-star rating, healthy sections rating, and staffing levels, both overall and for RNs.

Convergent validity testing was conducted and the correlation results showed negative correlation between facility level weight loss score and the overall quality rating, healthy section rating, and RN staffing.

Some reviewers did voice concerns with the

convergent validity correlation results, citing weak negative correlations between the facility level weight loss Q1 score and the overall quality reading.

One reviewer noted that although low correlations are common, these are lower than what is typically scene, indicating that overall nursing home quality and staffing may have little impact on residents likely losing weight. Meaning, weight loss is more due to patient conditions, and that a nursing home may have less control over this instead of the quality of care provided.

Seasonal variation was also tested and showed highest weight loss in Q1, with progressively lower rates in Q2 through Q4.

Methods panel reviewers also list concerns with the developer's decision not to risk adjust. The developer explored risk adjustment but stated that their attempts to develop this model were unsuccessful and resulted in a low R-squared value.

A reviewer noted that this might reflect the tight range of scores on this measure, leading to questions about its relevance. Because if there are not specific risk factors that could lead to weight loss and be addressed through interventions, it's an inappropriate quality measure.

In addition, reviewers noted that the literature indicates there are potentially addressable risk factors for unintentional weight loss. And reviewers would have liked to see which co-variants were tested in the model that had no predictive power, and since they were surprised that none of these factors were associated with the weight loss.

I will hand this discussion over to the lead discussant for this measure, who I believe is Jeff Geppert, to lead the committee in a discussion of validity.

Member Geppert: Thank you, Tamara.

Good day, everybody. So, first, thank you to the measure developer for preparing what I think overall was a very carefully prepared submission document, and for responding to the additional questions raised by the review panel. And, of course, thank you to the NQF staff for so excellently preparing all the materials for review.

So, the issue here really is kind of a borderline, as you can tell from the voting results, about whether the empirical validity results related to the performance score, you know, do in fact meet this threshold of being, you know, both satisfactory, you know, and sort of appropriate methodologically, you know, to demonstrate, you know, validity.

And so, part of the issue, I think, is that when the measure submission form presents the measure validity results and the methodologies used to assess validity, there's not really a very strong sort of assertion made about what the validity tester intended to show and sort of why we would expect them to show it.

So, and the results are so a little bit sort of ambiguous about that, how we're supposed to actually interpret, you know, the results that are, that are presented to us.

So, as Tamara mentioned, there were basically five different demonstrations of validity. One was a converging or construct validity where there was an hypothesized sort of similar care processes that underlied -- underlay different outcome measures. And so, one would expect them to be correlated at the facility level. And the results did in fact have positive correlations where those were anticipated, and negative correlations where those were anticipated.

But as was mentioned, the correlations were very low, sort of very weak, not really terribly compelling.

There was a demonstration of variation across states, but it wasn't really clear what this was supposed to tell us. There was some discussion about how if there is not variation across states, then that means that state payment policies and demographics were not sort of causally related to any sort of variation. So, it's kind of a negative inference that we were supposed to draw.

And, similarly, around seasonality, if there was not variation across seasons then that was not, again, sort of the causal reason why we might see any variation, although they note that this is a 4-quarter rolling average, so you wouldn't actually expect seasonality to be sort of the causal reason why we see variation.

There was a stability analysis presented. And, again, it wasn't exactly clear what we were expecting to see. Were we expecting this to be stable because there's a strong reason why structural elements of the facilities are driving performance so we would expect that, you know, stability to persist? Or are we expecting there not to be a lot of stability because there's sort of reason to expect that people are improving performance?

It's not clear what inferences we're supposed to draw from that.

And then there's kind of a confidence interval, you know, how many are above and how many are below the national means. But you get sort of an indirect, you know, if there are, if there's variation in an outcome then that sort of infers that there must be variation in process.

So, my low vote was basically because I felt like the results that were presented were they're not very compelling empirically. And the methods really didn't support the inference that we need to make which, you know, the essence of it is are there things that better-performing facilities are doing or that they have that worse-performing facilities are

not doing or do not have that is specific to this measure's focus.

And there's really nothing in the empirical results that were presented that relate specifically to things that they're doing or have that are related to this specific measure. And so that was sort of the reasons sort of for my low vote on this particular measure.

It seems like, you know, the real validity question here is whether performance on this measure is due to something very structural about the facilities that performed high or low that could have repercussions across the wide spectrum of outcomes. Or, again, is there something specific, some specific processes that are related to weight loss prevention, you know, that are really driving the results. And so that's why I felt like the methods weren't really designed to inform that question very specifically.

So, then I'll just sort of stop by talking about the risk. I think the concerns about the risk adjustment basically fall into that sort of the same line that, you know, if this is really about something that the providers are doing to prevent weight loss, then one would expect to see greater weight loss in populations that are highly at risk for that. And the fact that there was no empirical data presented, you know, that shows that, again makes one wonder what it is that this measure is actually showing.

And so I'll just end by saying this is probably a very important metric for the program to track, but I think the question for the panel is do the results presented and do the methods applied, you know, really support the use of this as a quality metric where there is some inference being made about the behavior at the facility level.

So, I'll stop there and see if my colleagues have anything to add.

Ms. Elliott: Great. So, we'd welcome comments from other subgroup members for this particular measure.

Co-Chair Teigland: I'll weigh in.

I totally agree with what Jeff said. I also voted this low. I also voted it low on liability for similar reasons. There's just very little variation in scoring, it didn't seem like they, you know, really could defend good versus bad.

You know, I find it hard to believe, having worked with the MDS for many, many years, I mean, weight loss clearly is important, that it's not associated with some of the, you know, co-variants that should be stated that apparently have no relationship: things like, like depression, like cancer, Parkinson's disease, cognitive impairment, eating dependency. You know, there's an item on the MDS that leaves 25 percent or more of their food uneaten.

That wasn't associated with weight loss?

I just -- We have chewing and swallowing problems, for example. But they didn't show any results of what was tested as co-variants in the model. And so, and, yeah, just the inability, the instability of the measure across facilities, the low variance in scores across facilities, lack of risk factors that seem to affect the outcome. I just have concerns that this measure is actually capturing are you doing a good job at preventing weight loss in these nursing home residents, so.

Ms. Elliott: Thank you, Christie.

I see a hand raised from Matt Austin.

Member Austin: Yes. So, I, I also found the measure score validity testing to be underwhelming for sure. But as I think we all noted, the data element validity testing was actually quite strong.

And so I was a little bit in a conundrum of how to

sort of synthesize those two together. So, I actually did consult the algorithm that we are asked to follow for making assessments around validity. And then following that it seemed to lead me to the sort of direction that we should be voting moderate for this.

That if we felt like the data element validity testing was strong and met the requirements, that that would earn it a score of moderate even if we found the measure score validity testing to be less than fantastic.

That's how I got to a score of moderate was following the algorithm.

Ms. Elliott: Great. Thanks, Matt.

Larry Glance, you have your hand raised.

Member Glance: Sure. Thanks.

So, I strongly agree with the comments that Jeff and Christie made. I think that it is very difficult to believe that certain case mix differences across nursing homes would not account for the differences in performance. Certainly there are many different medical conditions that would be associated with more weight loss.

And the fact that the measure developers did not provide any details in terms of the exploratory analysis that they made to me really made a pretty strong argument to grade this as inadequate risk adjustment, and being a very significant threat to validity. And that was why I voted low on validity.

Thanks.

Ms. Elliott: Thank you.

Okay. At this point I do not see any other hand raised or comments from the subgroup. So, the next step in our process is to allow the developers 15 minutes for an initial response, and may respond

to the SMP questions that were, and comments that were raised.

Who do we have on the call for this measure?

Dr. Nagavarapu: Hi. This Sri Nagavarapu from Acumen, and Chang Liu from Acumen developer is also on the call.

Chang, are you able to speak okay?

Ms. Liu: Yes. This is Chang from Acumen. Can you guys hear me?

Ms. Elliott: Yes, we can hear you.

Ms. Liu: Great. Okay, I guess I'll start.

Thank you very much for the introduction and the discussion. And that may address some of the concerns with risk adjustment. In this intro, while reviewing the preliminary analysis we noticed that most of the comments in the discussion guidelines are incorrectly referring to results from 2015. So, given that, I may upgrade some of the task assignments to make sure we're speaking to the same results.

The discussion guide describes that we are only correlating weight loss with global measures star rating. And we would like to know that we did include specific quality measures on convergent that we have seen. And there is a correlation around .1 in the expected direction when comparing with functionality.

In addition, we also tested correlation with prevalence of pressure -- and find that's a positive correlation of .15. This suggests that this measure is indicative of better care quality when comparing with other endorsed measures.

Even though these correlations are not high, we found them meaningful. We don't usually expect nursing home quality measures to have high

correlation. And when designing these nursing home measures, CMS intentionally selected different measurement areas to avoid redundancy. And, in fact, this level of correlation we observed here is in the typical neighborhood of correlation in other NQF endorsed measures that chart with major injuries, which was endorsed in the previous cycle, UTI and catheters which are endorsed the year before.

And the main issue of discussion I've heard so far is the insufficiency for risk adjustment. We were able to conduct passing analysis in response to the comments, and specifically address some of the variables, Christie, you raised. However, our passing results suggest that this measure should be maintained without risk adjustment.

So, let me go to these. First, we ran a logistics regression model using Alzheimer's, dementia, and depression. The predictive power of the model is extremely low. The C-statistic is only .51, which is pretty much the same as a coin toss. This risk adjustment model has minimal impact on provider performance remedy was 97 percent of providers. They were in the same decile before and after risk adjustment.

We also considered some additional risk factors the SMP recommended, including cancer, swallowing disorder, and dietary-related items.

However, we have compensated for these items. For cancer we have stated in our developer response but there is a data collection issue.

And for dietary-related items, for example, mechanically-altered diet, our concern is that these are service provision items under the control of the facility, so it may not be appropriate for risk adjustment.

In addition, mechanically-altered diets in swallowing disorders are both items, including eating dependency, are both items used by the PPS

payment model, which may be adopted by many states in the near future. Now, this may introduce sustainable effect in the practice of coding swallowing disorders, which will make it inappropriate to include as a risk factor.

We have this concern because since the implementation of the PDPM model about two years ago, we have observed that underquoting has surged from 4 percent to over 17 percent among this mixed population. And now, starting from next October, over 30 states started to collect information to evaluate whether they're going to transition their state payment to PDPM.

So, we have this concern that once a transition to PDPM that it's possible that we'll see this similar surge among the states who enroll in PDPM at their state clinics.

So, regardless of these concerns, we did run a model using these variables on swallowing disorder, diet, and cancer, and observed that the impact on provider ranking is quite limited. 80 percent of providers stay in the same decile. So, given this limited impact in the concerns mentioned above, we think it is premature to risk adjust this measure, and choose not to risk adjust this one

This concludes my initial response. And thank you very much for giving me this extra time. I'm happy to speak more to some of the points I touched in the upcoming discussion.

Ms. Elliott: Great. Thank you so much. We appreciate your comments.

So, at this point we can open discussion to the full SMP and proceed to discuss this measure.

Jack, you've placed a comment in the chat. Would you be able to share your question and comment?

Member Needleman: Sure. It's a comment, not a

question.

I was just struck by Jeff's phrasing of what's going on with this measure in terms of steps people, the homes, can take to avoid weight loss. And I actually think the measure was introduced, and the concern about this and pressure ulcers is about poor care, not appropriate care.

And, you know, I'm old enough to remember the New York State nursing home scandals where some facilities were serving patients oatmeal three times a day. So, I would not be looking for positive treatment here.

So, that was one comment.

The other one was is when I said the floor for the moment, is I think the issue with the risk adjustment is it's like the Hound of the Baskervilles, it's the dog that didn't bite -- didn't bark. We're expecting things about the patients to influence their weight loss. And you're not, and, Acumen, you're not seeing it in your analysis. And I think that's making people nervous and concerned.

And I'm wondering if you have reflected on that as you've looked at your risk adjustment results, and have any thoughts about why things that ought to be affecting weight loss at the patient level don't seem to be in your risk adjustment model.

Dr. Nagavarapu: I'd be happy to respond to that if there's a good time.

Ms. Elliott: Yes. Go ahead.

Dr. Nagavarapu: Great. Thanks for the chance to respond here.

So, in the risk adjustment results that Chang referred to that we did in response to seeing some of the points that you all have raised, you do in fact see that at the patient level there is an impact in the direction that you would expect for these co-

variants.

So, for Alzheimer's, dementia, and depression, the odds ratios are all positive and significant. So, the odds go up if you have Alzheimer's, dementia, and depression for weight loss. But the effects are small.

And as Chang noted, the predictive value overall is limited, suggesting that there are other factors going on. And there may be compensatory responses from facilities when they notice that a resident has one of these features.

The other results that Chang notes is about mechanically-altered diet and no swallowing disorder, those co-variants. Those results are extremely strongly predictive. So, the odds ratio for mechanically-altered diet is 1.61, and significant. The odd ratio for no swallowing disorder is, as you would expect, less than 1.0 at .47, and significant.

Again, there are conceptual concerns that Chang noted for reasons that CMS would not necessarily want to address for these co-variants because CMS has observed increases in coding of these items after the introduction of the patient-driven payment model. And so, risk adjusting for these items has a serious implication that we're trying to evaluate, and we'll have more evidence on over the coming years as states adopt the payment model.

So, the thing I want to make clear is that these co-variants are predictive in the way that you would expect. It's just that some of them, like Alzheimer's, dementia, and depression are fair -- seem to be fairly uniformly distributed across facilities, in which case they're not having a large impact on measure scores.

And then for the ones that have really strong effects, there's concerns about both coding and service utilization that are driven by the payment model in terms of whether or not we want to risk adjust for them. For that reason, you know, we

wanted to show you these results, kind of talk through them. And then, also, think through sort of the underlying mechanisms for what can happen.

And I think, as you noted, Dr. Needleman, really like the measure's designed to sort of highlight poor performance in these areas. And we think that the measure variation, if you look at the inner quartile range, the 10th to 90th percentile, is fairly dramatic and is able to satisfy that.

The correlations with the staffing ratings as well as the other quality measures are also consistent with those underlying processes having an impact.

So, I'll stop there and see if there are any questions.

Oh, and I see in the chat, we agree on the algorithm. To the extent that data elements are valid, we think that the algorithm also implies a moderate score.

All right. But thanks for the opportunity to respond to those questions.

Ms. Elliott: Okay, thank you.

Other comments or questions from the committee?

Co-Chair Teigland: I just might weigh in there. It also seems a little impossible maybe. You know, 20 percent of facilities changed scores by three dec -- by more than three deciles, three deciles or more from quarter to quarter. That seems sort of random as opposed to really quality of care issues, which might be why you're not finding the correlations with all the things that weight loss should be correlated with, even staffing levels, even over on the single quality, you know, even over 30 percent, you know, changed deciles, one to three deciles.

So, there's very little instability in the measure that led to all these factors were low. So, I just think that I still feel uncomfortable with that.

Dr. Nagavarapu: Thanks, Dr. Teigland.

One thing we wanted to note on developer response is we provided some information, more information on the measure score reliability. And I want to emphasize that the average reliability is .76, with a median of .78, which is quite high. Even the 25th percentile is at .68, which is close to the .7 standard that many use for moderate to high reliability.

Co-Chair Teigland: Which reliability are you talking about? Because I know the SNRs was only 0.078, which is very low.

Dr. Nagavarapu: The signal-to-noise reliability has an average of .76. And the 50th percentile is .78, which is, which is very high.

I am not sure whether the old 2015 testing form has another number that you have in mind. But the submission refers to the average score. And I think there was some questions about the distribution. And we provided that in the developer response.

Thanks.

Co-Chair Teigland: Okay.

Ms. Elliott: Okay. Checking for hands raised.

Okay. In the chat there's been a request to -- hold on a second. Let me find it again.

There's a request to briefly surface the algorithm. And the comment, Matt Austin, you had put a comment in the chat stating that your impression was that the algorithm should guide our decision making. That's not to say that I agree with the algorithm, but it at least provides a standard evaluation process for all measures and all reviewers.

So, Alex, did you want to comment more on that?

Member Sox-Harris: Sorry, I'm not in a great audio-

visual space here.

But, yeah, I just wanted to look at the algorithm. I could try to find it in my files, but it might be helpful since that seems to be an important aspect of this measure evaluation, whether data element validity prompts in some sense our concerns about opportunity for reliability.

I just would benefit from reviewing that quickly.

Member Geppert: So, just a clarifying question about that. The data element evidence if from the original RAND work from 2009 and 2012, which I'm sure was very rigorously done.

Does that really meet the criteria that the testing results sort of reflect in the data that's actually used? And then is there no shelf life, you know, on those kinds of investigations?

Dr. Nagavarapu: That's a great question.

Chang, would you like to talk about the stability of the MDS assessment at this time?

Ms. Liu: Sorry, I had myself double muted.

Yeah, sure.

So, there was, even though we know that the RAND study was already conducted a while ago, one thing we want to point out is that there is a follow-up study in 2012 that basically confirmed everything they have done previously as being valid.

And on our end, we checked the MDS item, the instructions, and figured out for the weight loss items there is no change in the item design.

And I think this data element validity is also supported by our measure changes. I want to note that for the stability analysis where we looked at the deciles in percentile ranges, we, I should not that in this analysis and the related passing form, only 7.5

percent of providers changed three or more deciles.

So, this is in support of the stability of the measure, it is pretty low. We will take a look at changes from one study period to another.

Ms. Elliott: I think we want to -- there are several questions about the algorithm that came up in the chat. So, before we move on, I wanted to see if we could pull up the algorithm slide. And we have Matt Pickering on the line who could speak to some of the elements of the algorithm.

Dr. Pickering: Thanks, Tricia.

Hi, everyone. It's great to see everyone again.

So, I think there was a question around prioritizing or maybe considering data elements in conjunction with the accountable entity level testing.

So, as you can see on your screen, it's algorithm number three. This is pulled from our measure evaluation criteria.

I wanted to just kind of note the flow of this, obviously, where at the very minimum, number one, there should be consideration of the potential threats to validity. So, those must be considered amongst everything else. Right? So, at a minimum we must consider that, that must be empirically addressed, and if so, moved to yes.

And then the blue boxes really talk about the empirical testing of validity at that patient level.

And then you go down to the yellow boxes as well, where you start to get into the accountable, accountable entity level.

If you scroll, go all the way to the right, you'll see the ultimate rating, the high, and the moderate, and the low. You also see these little symbols there, rate as high or lower with the little two plus signs, the same thing for moderate.

If you go down to the bottom -- and I'm not sure we can zoom in -- but what it basically references is that the overall rating you can get, the highest rating you can get is high for validity testing. But this also may be lower, depending on the strengths of the patient accountability level validity testing.

So, if that patient accountability level, validity testing is provided, it must be considered. So, if there is that accountable entity level that is provided, and as well as the patient level data element type of testing, then that patient level data element should be considered as well.

And this may alter the rating for the accountable entity level which you sort of go back to. If the patient level is sufficient, you may, you may find that's more of a moderate rating as opposed to the high, which is the highest achievable rating you can get with the accountable entity level.

Does that answer your question for the SMP committee or group? I apologize. Matt Austin, specifically, I think you raised that question.

Member Austin: Yes. I mean, for me I, the way I thought about this is I walked through the blue steps first. And answered yes that testing -- and I can't see the blue part. But, you know, I wound up in the yellow which then, you know, I did not think the results were, you know, really very demonstrable of validity, so I answered no to that. Which then took me to the orange which eventually got me to moderate.

So, I mean, you -- I can appreciate Jeff and others' concerns about you don't necessarily buy in that the data element validity is (audio interference) and obviously we're going to reach a different conclusion.

But if one does find that to be sufficient, then we are stuck in this situation of having to evaluate both and, at least for me the algorithm seemed to guide

me toward that. One (audio interference) moderate.

And, once again, I'm not necessarily defending this particular algorithm. I'm just saying that this is the algorithm that we've been given to use for evaluations this time.

Ms. Elliott: Great.

There's a hand raised from Sherrie Kaplan. Sherrie, did you have a question or a comment?

Member Kaplan: Well, this is kind of an old saw with me, but there's no difference between high and moderate in terms of discrimination for pass and no pass. And that's been a concern of mine all along. There's no meaningful difference the way NQF evaluates these measures between moderate and high. So, it's somewhat of a moot point.

I mean, we're spending time on this issue where there's not going to be any difference in the ultimate outcome of the vote to rate it moderate or high. So, unless there's something else we need to -- you know, if it's moderate it's still going to pass. So, I was still worried about NQF's ultimate distinction here. But, maybe this algorithm could be revised.

Co-Chair Nerenz: I wondered, David, if I could just ask a question on the algorithm and so, following up on that. Unfortunately, it's either my eyes or the screen so I couldn't read anything that's on there.

I think this situation we face is that the rules of the game require either patient level or entity level data. This could be true of either reliability or validity. The developers give us both; and one is good and one is bad.

The question is, does the algorithm tell us that one level trumps another level, or one comes first and one comes second? Or does it tell us that if both are provided, we have to consider both in our

evaluation? That's the question I would ask about the algorithm. We're going to see this several times today.

Dr. Nagavarapu: And just to throw one last piece of information in.

In terms of the measure level validity testing that we did, I think some of the comments that have been made so far are referring to the 2015 testing form. So, like, the specific numbers that Dr. Teigland referred to refer to the 2015 testing form which, among other things, uses a quarterly measure which is different than the annual measure that CMS has moved to for public reporting.

So, I would strongly encourage people to use our submission rather than the 2015 testing form when we're making any decisions here. And as much as possible, we've been trying to correct that along the way. But several of the comments so far have been referring to the 2015 testing form where, naturally, stability is lower because the measure is different.

So, please make sure that we're referring to our submission.

Thanks so much.

Ms. Elliott: Matt Austin, did you have another question or is that a leftover hand raise?

Member Austin: Yeah, I was just going to comment real quick.

Ms. Elliott: Okay. Sure.

Member Austin: I mean, to me the distinction that I'm trying to make in my mind is between medium and low, which has, you know, real implications for measure development.

I would agree that high and medium are equivalent for purposes of this conversation. For me, the conversation is just saying is it a medium or is it a

low rating that should be assigned.

Ms. Elliott: Thank you, Matt.

Larry Glance.

Member Glance: All right, thanks.

So, I always find this algorithm incredibly complex and difficult to work through. I don't know what other people think about this. But when it comes to threats to validity, whether on this risk adjustment or others as well, for example, what we choose to exclude and not to exclude, those threats to validity are they threats to overall validity, or are they threats to accountable level, or measure level, score level validity?

I've always interpreted them as threats to overall validity. And so, in my mind when I think through the algorithm, I think that if the risk adjustment is lacking or poor, then that is a threat to the overall validity of the measure, and then that would determine my decision as to whether or not to do a high, moderate, or low.

I don't know what other people think about that.

Dr. Pickering: So, Larry, this is Matt. And thanks for the question. And you are correct is that this is why it's sort of at the very top of that algorithm it, like, has the measure or developer explored empirically the potential threats to validity? So, those should be at least empirically assessed at the base of validity considerations. That algorithm starts there. And if that has been empirically assessed, then you start to move to the actual testing of the other components, whether it be the data element or the accountability level.

But if there is both that have been presented, both should be considered. And so, if this is a case where the entity level, accountable entity level had a high rating but the patients, the patient level maybe did

not look so good, you may change your vote from high to moderate because both need to be considered.

If one is, if the accountable entity level is not that great but the patient level looks good, the highest possible rating for a patient level only is moderate. But both should be considered. And if the highest rating is high, validity, validity for the accountability level but the patient level may not be great, you're still considering that. But you may shift the decision making high for the accountable entity to moderate because the patient level validity testing is not great, is not up to what you would find to be acceptable.

So that's sort of how it's framed within our current algorithm is that if both are provided, both should be considered. And that may change the ratings for one or the other, depending on the results of each.

So, again, in this case I think if the accountable entity level is not looking great or sufficient to the standing committee or, excuse me, to the SMB, then looking at the patient level and determining whether or not that is sufficient enough to move forward with a moderate rating or something else. But if it's just patient level it would be the highest rating would be moderate.

Ms. Elliott: Sherrie, did you have a comment or a question?

Member Kaplan: I think this is back to Matt Pickering.

I know that in the algorithm itself the votes for the high and moderate in the validity zones are flagged as "or lower." So, that means that that's the highest possible rating you can give it. You can give it a lower rating, but that is the highest possible rating you can give it.

And the second thing is, is we extract the data

accurately from the data source, whatever it is, medical record or questionnaire or whatever, but at the accountable -- and this, I think, is framed somewhere in our discussions, and I don't see it easily in the algorithm -- when a measure sees a certain level of accountability, it must be tested at that level.

Now, if it's tested and it doesn't make the standard, you know, just because you can get it accurately from the data set doesn't mean that then it should be discriminating between facilities that that's what it's going to be used for.

So, I'm still of concern that we're struggling with something here that the algorithm isn't clear on, and it's leaving some of us confused about, yeah, well, we did it well at the patient level but it's not discriminating between facilities, so maybe it shouldn't be used in that way.

Ms. Elliott: Thank you.

At this point --

Dr. Nagavarapu: Oh, and just quickly in terms of the point about discriminating between facilities, we definitely want to emphasize that our submission shows the reliability is at .76, which is 10 times the number that Dr. Pickering was citing earlier of .07.

So, by historical NQF standards, the measure does an excellent job at discriminating between facilities in terms of signal-to-noise metrics and ICCs.

Thanks.

Ms. Elliott: Thank you. So, at this point we have exceeded the time allotment for this discussion, but we did include some additional information regarding the algorithm.

So, Christie, could I defer to you if we're ready to move to a vote on this measure?

Co-Chair Teigland: Yeah. I think, I think so.

So, we're just voting on validity here. And I did go back, and there was additional information in the submission form. I think I probably was looking at the testing form. That was confusing probably for many of us, I think. I'm not sure why that testing form was shared with us, but we probably should have just had the new, new data if that was all updated and not even used anymore.

So, sorry for the confusion there.

I think some of the issue, though, about validity that we've discussed, you know, some of us still may not be satisfied that we're not seeing correlations with things that we think it should be correlated with. But, yeah, I think as far as I'm concerned we can go ahead and vote.

Ms. Elliott: Okay, great.

So, Hannah, can I turn things over to you for the voting?

Ms. Ingber: Yes, certainly.

So, voting is now open for validity of Measure 0689. Your options are high, moderate, low, or insufficient.

Member Warholak: Hi. This is Terri.

I'm not really sure where we get the link to vote. Can you guide me through? Thank you.

Ms. Ingber: We're using the same link as was emailed for the test votes.

Member Warholak: Ah. Okay.

Ms. Ingber: Do you have access to that, Terri?

Member Warholak: Yes, I do. Thanks.

Ms. Ingber: Okay, great.

And only the subgroup is voting on this measure, only the Subgroup 2 that reviewed this measure is voting.

(Voting.)

Ms. Ingber: All right, just confirming the results.

Member Warholak: Okay, I lied. I cannot find that link. Can you resend it to me, please. Thank you.

Ms. Ingber: Yes. Just one moment.

(Pause.)

Ms. Ingber: All right. We do have quorum for this voting. The minimum is eight, and we have ten results. So, I'm going to share the results.

So, voting is now closed for Measure 0689 on validity.

We have 0 votes for high; 2 votes for moderate; 8 votes for low; and 0 votes for insufficient. Therefore, the measure does not pass on validity.

Thank you, everyone.

Ms. Elliott: Thank you, Hannah.

The next item on the agenda is our break. We have it scheduled for 25 minutes. So, we will reconvene at 1:30 p.m. Eastern Time. And we will keep the discussion going.

The measure up for discussion when we come back from break is 3649e.

And with that, do we have a slide? And we'll stop the recording during break.

So, the slide says 1:00 p.m. We're a little off schedule, so we'll be reconvening at 1:30 p.m.

Thank you, everyone.

(Whereupon, the above-entitled matter went off the record at 1:05 p.m. and resumed at 1:32 p.m.)

Ms. Elliott: Okay. As we head into our next section we'll be starting our discussion on measure number 3649 within the surgery area. And LeeAnn White, one of our NQF directors, will be providing background on the measure. LeeAnn?

Ms. White: Thank you, Tricia. I just want to start off by asking if you can hear me loud and clear? Okay, wonderful. So good afternoon everyone to the Scientific Methods Panel, and all in attendance today. I hope you all had a pleasant break. My name is LeeAnn White and it is my pleasure to support the surgery project team and to introduce our first surgical measure, which is 3649-E, which is the Risk Standardized Complication Rate following Elective Primary Total Hip Arthroplasty, or THA, and, or, Total Knee Arthroplasty, which is TKA.

This is a new measure for the fall 2021 cycle and the measure developer is Brigham and Women's Hospital. This electronic clinical quality measure, or eCQM, quantifies the risk standardized complication rates following elective primary THA and, or, TKA at the clinician group level for adults 18 years and older across all pairs. The rate is expressed as a percentage where a lower rate is indicative of higher quality care. The outcome is defined as any specified complications occurring from the date of index admission to 90 days following discharge.

This outcome measure is analyzed at the clinician group practice level and is intended for ambulatory care, inpatient hospital and outpatient care settings. The type of score is rate proportion and for risk adjustment this measure uses a statistical risk model with ten risk factors.

So moving on to reliability, this measure received a consensus not reached rating after initial subgroup evaluation. A developer conducted reliability testing at both the patient encounter level and the

accountable entity level. The developer used the feasibility score card for patient encounter level testing to assess EHR data availability, accuracy, terminology standards and workflow. All 23 data elements received a score of 1 out of 1 for both Cerner and Epic sites.

In regards to concerns related to patient encounter reliability testing, one SMP member questioned how comparing socio-demographic characteristics of patients in both the random tests and validation samples demonstrated element-level reliability. At the accountability level, entity level, the developer used the test-retest approach comparing the agreement across clinician groups on the performance measure. The predicted expected ratios ranged from 0.719 to 1.404 with a Spearman ranked correlation of 0.978.

For variability across clinician groups, the intraclass correlation coefficient is 0.006. The subgroup members raised several concerns during the initial evaluation of reliability testing. One member raised the question of the measurement's performance period for a clinician group who will use and report this measure, and that the impression would be the measure's window is one year from January to December. Yet if a testing used four years of data and each random sample -- testing and validation -- contains 50 percent of the data, then each split half sample might contain approximately twice as many records compared to the sample size available in the actual implementation in a single calendar year. The member notes that is in the case, then the reported Spearman correlation coefficient might overestimate the correlation obtained across two calendar years in the actual implementation.

SMP members also raised a couple of concerns with the intraclass correlation's wide confidence interval around the ICC estimate. And then also, they -- the small number of practices, the skew in distribution

of patient characteristics between practices and lower portions at the Cerner site. Lastly, the developer assessed the risk adjustment's logistical regression model, calculating the model strength in predicting a complication event. ICC statistics calculated was 0.672.

Validity was tested at both the patient encounter and the accountable entity levels, achieving a moderate rating for validity, following initial evaluations from the science subgroup. The developer conducted a validity testing of the data elements through manual chart abstraction of the numerator, denominator, and exclusion data to the eCQM calculation noting any disagreements between the EHR and the eCQM. The kappa coefficient for this agreement ranged from 0.8333 to 0.9495.

The developers assessed face validity at the accountable entity level through a seven-member expert panel process. The TEP was engaged throughout the electronic measure development process, providing feedback during specific points, and the final measure specifications were provided. Eighty-five percent of the TEP members agreed that this measure could be used as it distinguished good from poor clinician group-level care polity related to patient safety.

And lastly, just a few concerns were brought forth by the subgroup SMP members during the initial evaluation. First, a concern was raised with the provided complication rates seen at the hospital level and not the clinician group level. Some SMP members raised concern with accountable entity testing, noting that the developer assessed only face validity, and that the TEP was asked only one question during face validity testing. And I will now hand it over to Sherrie Kaplan to open up discussions around some of the concerns related to reliability and validity. Thank you.

Member Kaplan: And I -- I don't have pretty much anything else to say. You did such a great job of summarizing all the issues. I think I'll just add a few more things. One is the concern about the ICCs at the clinician level -- even though it's a small sample size, and the developer says that when the larger sample size is included the limited variation they saw is going to expand and they're counting on that -- it raises a lot of concerns because it's less than one percent of the -- of the variation in this measure attributable to clinical groups, as data suggests. So that raises some real concerns for me, especially since the -- some of the other issues come up.

Another issue that I don't think you raised was the response rates are pretty low. At the clinician level it was 18.5 percent. At the group level is 32.3 percent. So again, you know, the generalizability of these results is of concern. There are ceiling effects in this measure -- 37 to 46 percent. And granted they -- they look at WOMAC and they look at the sort of ceiling effects there. Then you've got -- you've got ceiling effects, that's a -- that's another problem when you're trying to discriminate -- if you've got limited variability, you've got to -- you know, discriminating between practices is of concern.

There's -- 50 percent of the sample was excluded due to -- about 37 percent were missing PROs in the sample. Ten percent were missing risk factors and an additional two percent were missing some kind of attribution to clinicians. So again, generalizability of responses are an issue.

And as you said, the -- there were 17 groups that are EHR sites and two different vendors, Epic versus Cerner, in this population. That's a lot -- that's a very small sample to kind of work with in terms of data.

And finally, I think, I have concerns -- well, the face

validity issue is there was a single question asked to seven members of the TEP. And that's the basis for face validity. So that was a concern. In the risk model there was concerns raised because, for example, there are some of these risk variables that are -- are confounded with site. So osteoarthritis was present in 99.3 percent of folks at MGB versus 28.5 percent of folks at Cerner sites. So that's a -- that's a concern when you've got variables like that. And the -- the -- the developer used ridge regression which suggests there might be colinearity among some of these risk factors but they didn't give us any collinearity statistics for -- you know, and when you've got 29 comorbid conditions, that raises some issues about potential for a risk model discrimination calibration.

So that's why my concerns led me to believe that the reliability was -- I gave it a score of low. And also I gave the validity scores low. But my colleagues were not in agreement on that one. So that's -- those were my concerns in a nutshell. And I'll leave it to some of the other folks on -- in Group One to comment.

Member Romano: This is Pat, can I speak up?

Ms. Elliott: Yes, go ahead Patrick.

Member Romano: Hello, yes, this is Patrick Romano. Yes, just a quick question of clarification for Sherrie. When you're referring to response rates, what are -- what are you referring to here? I don't think I saw those.

Member Kaplan: The exclusions -- and it was included in exclusions. So two -- the total number of folks that weren't included in the sample were because they were missing the -- the HOOS and KOOS of about 37 percent were those. Another 10 percent were risking the risk factors and 2 percent were attributable to the --

Member Romano: Okay, okay. Yes, I think that --

what's our process here? Are we going to discuss reliability first and then validity? Or are we discussing them together?

Member Kaplan: I didn't think we were discussing validity at all. I just raised my concerns because I was asked about my concerns about giving it a low score so I wanted to enumerate what those were -- especially for the risk model confines with site.

Member Romano: Yes. I think the measure was pulled for a discussion of both. So are we discussing reliability first and then validity? Or are we discussing them together?

Ms. Elliott: Let's focus on reliability first because that was consensus not reached.

Member Romano: Okay, so yes I -- I think it's an interesting question which a number of the measures from this group illustrate, which is what do we do when there's a striking difference between two approaches to assessing the reliability at the accountable entity level? So one approach is a split-half reliability approach which simply looks at the random error, if you will, due to splitting the -- the sample and comparing the performance estimates from the two split halves.

The other approach actually looks at the -- at the variability that's attributable to the accountable entity level across all the variability using an ICC approach. And these approaches appear to give markedly different estimates. So I am interested in what my colleagues think of this situation and how we should approach it.

Member O'Brien: This is Sean, if I can -- I'll jump in at another Group One measure. I think LeeAnn gave a great summary and a great -- various comments as well -- and Patrick's now. Things I would add is that when I look at the results of the correlation analysis -- the split-half correlation analysis that saw results correlation of 0.97, this comment I

made I think maybe didn't make it into the development -- the discussion guide, but I -- it made me wonder about an error because if you -- in the table that was provided, I think it may be Table A -- if I understood correctly, they gave kind of the -- the entity-level results for each estimate in each split half along with a confidence interval around each of those estimates. The confidence intervals around each entity's estimate were quite wide, yet they happened to correlate perfectly across two different samples. And that's -- it struck me as being potentially inconsistent -- that you wouldn't expect such a high degree of correlation if there was so much uncertainty in the point estimates.

Another comment I would make -- one, I think it could be in the developer's interest to clarify what -- with the ICC estimate that's being reported is describing a -- a metric that's applicable to the data once you aggregate across all of the patients contributing to the measure or describing variation in a single measure. Because one of those we'd expect to be an extremely low measure, and I -- it wouldn't think much of it. But if it's actually describing, you know, the -- the -- the measure of ICC after aggregating across bases within a site, then that -- that's an extremely low number.

Regardless of either interpretation, I notice that the confidence interval around the ICC extended all the way down to below zero. So you could -- based on the data, you couldn't rule out the hypothesis that there's no -- between no true signal variation between the entities -- the -- those factors contributed to my low rating for reliability.

Member Needleman: I've got my hand up, if I can comment on this reliability issue.

Ms. Elliott: Yes, go ahead Jack.

Member Needleman: This is Jack. Yes, so I think this measure -- and we'll see it in some of the other measures that are presented here -- was tested in a

very limited number of sites. And those sites had standards of care and practice that seem to be very similar across the sites. So where -- where -- and the -- the developer acknowledges that. They're not seeing a lot of variation across sites.

So many of our site-level measures of how well does this discriminate against -- across sites given the small numbers are -- are going to be very poor. And this comes back to the issue that Sherrie raised at the beginning of the conversation, which is do we accept the -- the measure level -- the individual level, patient level reliability as well enough established and say go out into the world with an NQF endorsement and get more data? Or do we say, please bring us more data before you ask for an NQF endorsement?

If it's the latter standard -- or the latter -- if the latter is the approach, then this -- this measure does not pass our reliability standards. We're not seeing enough variation. We're not seeing it across sites to assess whether this measure accurately discriminates against sites. A -- a -- I'm prepared to acknowledge the reliability at the measure level is okay. But we have not got data or evidence here that it performs well at the entity level. And the standards are slightly different for this kind of measure than the -- you know, an instrument-level measure which we'll be looking at later.

But I think the -- the core issue here is not so much the statistics, but whether the -- ignore the fact that we're not seeing a variation, we've got a small sample in which we've tested the -- the data, please give us an endorsement, is going to be responded to positively or not.

Ms. Elliott: Thank you. Sherrie, I see your hand is raised?

Member Kaplan: Yes, I mean just to follow up on what Jack said. I mean, to me when new measures come in -- and this is for the NQF team. When new

measures come in like this, and it's costly to collect data. You know, you have to go out and collect and analyze data. Is there -- and this is probably not for discussion, but is there a provisional acceptance or something that you can do along the lines of what Jack just said? Give -- give a preliminary or some kind of early endorsement for new measures like this that are not new at the patient level -- this -- HOOS and KOOS have been around for a long time. They've got established patient-level reliability across multiple different studies, yada, yada, yada. But now it's being used at a different level which changes who you approach the reliability testing, et cetera, which again -- you know, you need a sample. You need sample. You need the ability to go test it.

So I am not sure what NQF endorsement does for the ability to garner resources to do that. But to me it makes a lot of sense for new measures who have given you some early data to get that kind of provisional endorsement. But that -- you know, that's maybe a conversation for a different time. But I -- I think this kind of measure calls for that and that's why I raised it early on.

Ms. Elliott: Yes, thank you Sherrie. Jennifer has a hand raised.

Member Perloff: Here we go -- off mute. I want to step back even further than where we've gone already in the reliability stage. This measure was -- the reliability testing was done in Epic and Cerner data and no other EHRs. Epic is about 25 percent of the market -- academic medical centers only. Cerner used to be 25 percent of the market, but that's 50 percent of the EHR market that has not been touched here. And those are lower-resource, simpler EHRs. So the fact that the reliability testing was done in such a limited set of EHRs is a huge concern for me because once this measure goes out into a broader universe, we have absolutely no clue that it will be reliable at all.

So to the point that Jack was raising about sort of how much data is enough, I think we're not even close on enough data. And this is a really major issue with all EHRs that we've received. But I have to raise that as a serious concern.

Ms. Elliott: Thank you for your comments. Any other subgroup members?

Co-Chair Nerenz: Yes, David -- I'm sorry that -- I'm finding myself at a loss. I go back to Patrick's comment -- are we even talking about the right measure here?

It seemed like there's a HOOS KOOS measure in this set, but it's not 36-49. Am I missing something here? I just went through the --

Member Kaplan: I just misspoke, Dave. That was just -- the HOOS and KOOS is the next measure. This is complication rates and it's attributable either to the hospital or the -- these are being attributable to the clinician groups. So these are complication rates following THA and -- and TKA.

Co-Chair Nerenz: Okay. I just didn't know how much of what we just said is relevant to this measure. And that's what I am trying to sort out.

Member Kaplan: I just misspoke. It -- all the rest of the discussion is relevant. The rest of --

(Simultaneous speaking.)

Co-Chair Nerenz: Thank you.

Member Kaplan: -- I just misspoke. I skipped over and went to the next measure inadvertently.

Co-Chair Nerenz: Okay, okay.

Member Romano: I just want to point out that technically the minimum requirement for testing in EHR systems, according to NQF, is simply more than one vendor. Developers who test on a number

of EHR systems, they feel appropriate. So as long as it's two, which we have in this case, that's probably not a grounds for not passing. Other things may be grounds for not passing, but not that in itself.

Member Perloff: I would extend my thinking -- you know, we wouldn't accept perhaps the testing of this measure if everybody was from sort of the same race, ethnic group or other forms of sort of bias. To me that's a serious bias that the patients and the EHRs included in this testing are from a very select subset of the healthcare delivery system, and it doesn't suggest to me that this measure would be reliable for a -- a provider group that can't afford those systems. So I appreciate that point, but it seems inadequate -- criteria.

Ms. Elliott: Daniel, you had your hand raised?

Member Deutscher: Yes, this is more of a question to NQF. So I -- I totally agree and appreciate the comments from my colleagues. But I'm trying to translate that given the NQF requirements. And a little bit following Patrick's comments. So as best as I understand, given the current requirements -- although there may be important issues with the representativeness of the data set used for -- for adjusting -- that may or may not be the case. But let's say it is the case. And it's clear that there is a -- there is a big issue with reliability at the accountable entity level. But given the current requirements, are these grounds for not passing the measure on reliability? I think they're not -- given my understanding of the NQF requirements. But I'd like a clarification on that if possible.

Ms. Elliott: Sure. Matt Pickering, could you address that? Did you hear the question?

Dr. Pickering: Sorry, no I did not hear the question. Daniel, can you repeat that?

Member Deutscher: Yes, sure. So let's -- let's assume there are issues with the representativeness

of the sample used for testing as -- as commented here. We also are pretty confident that there is very low variability between providers and the accountable entity level reliability is an issue. But given the current NQF requirements, are these grounds for not passing the measure on reliability? Or on the other hand, maybe just giving a moderate rating?

Dr. Pickering: So -- I think the question being the concern that the -- the sample is not generalizable? And that's the main concern?

Member Deutscher: Well, it's one concern that has been raised here and -- but I -- I don't see -- and I agree, may be a concern. My question is, is there a criteria -- a criterion about that, given the NQF requirements?

Dr. Pickering: So I -- I think the -- the data sample is something for the -- the standing committee to consider and discuss. If it is a -- sort of a concern with -- related to generalizability. However, you know, we referred to the algorithm within the criteria that points you to your decision making -- whether it be at the accountable entity level or the patient -- the patient level. That algorithm doesn't necessarily note generalizability, but it is something within the specifications of the measure that should be discussed and considered. Whether the -- the measure specification of the data source that's being used is appropriate. And especially for the -- the level that's -- that the measure is intended for and how the testing supports that and making sure that it aligns with that -- that level of analysis. So I -- I don't know if that really gives you a fine -- a definite answer, Daniel, but really just referring back to our algorithm within the criteria of how to adequately -- how to adequately come to a decision on the -- on the reliability testing.

Member Deutscher: Yes, okay thank you.

Dr. Pickering: And then, you know, the team also

just reminding that it -- you know, the -- we are not that prescriptive when it comes to low volume with -- with related to sort of data and sample sizes. However, it is something to where it should be considered and discussed by the SMP with the information that's been presented to you.

Member Deutscher: Okay, thank you.

Ms. Elliott: Yes, thank you Matt. Sherrie, you had your hand raised -- and then Dave.

Member Kaplan: Yes, I -- the -- the -- there is, from the data presented, including the inter-class correlation coefficients of less than 1 percent. The data provided suggests that there is more within- than between-practice variability. And when you look at the level of osteoarthritis in the risk model as -- you know, at -- at MGB, the 99.3 percent and the Cerner site's 28.5 percent -- that suggests that there is a lot of within-practice variability if you're combining those two groups. So it might -- you know, this speaks back to, if you had a different or a broader sample size, you might get some -- some more between- than within-practice variability. But when you've got that kind of askew, the -- it is concerning that there are this -- this -- that's what's driving inter-class correlation coefficient. That the -- the within-practice variability of some of these kinds of characteristics that could influence complication rates.

So my -- my concern is that you're shifting from a hospital-based complication rate to a clinician group complication rate. And then where is the appropriate attribution? And -- and the data provided suggests it might not be at the clinician level because I think it's premature to make that call without some additional information. So I'm kind of still stuck in this. You want this measure to kind of look for, but combine to a ceiling effect issues and other kinds of things that are going on. You've got the -- you've got a problem in the

generalizability of these data that may limit our ability to actually see a signal that's actually there. So I'm kind of on the -- given the data provided and the guidance provided, I'm still in a position where I'm probably on a low reliability. But I don't know how to deal with this provisional business about -- it's small and it's not -- it's limited data.

Ms. Elliott: David?

Co-Chair Nerenz: Yes, just a basic question either to my statistician colleagues or to the developer, when we get to that stage. In assessing entity-level reliability, there are two statistics presented to us. There's a rank-order correlation based on split-half analysis -- very, very high. There's an ICC value. It's very low. Essentially zero. And what do you do with that?

I had come into this with the idea that the ICC is perhaps the preferred or more informed of statistic, but I would be happy to be corrected on that. And the question is, if they're equally informative, do you subjectively average them in some way? Is ICC truly more informative than the rank-order correlation? Am I wrong on that? Some guidance on that would be helpful because the -- the ICC evaluated most heavily, in my judgment.

Ms. Elliott: Okay. Let's see -- okay. Larry, I see you have a hand up, but you're in the larger group, not in the subgroup. So I am going to have you hold your question just for a minute to give some time to the developer to respond. And then we'll circle back. Who is on the line from the developer, please?

Dr. Dykes: This is Patricia Dykes from Brigham and Women's Hospital. I'm here -- I led the development for these -- this set of eCQMs. I also have Stuart Lipsitz, our biostatistician so he can talk about the statistical issues. I just wanted to point out one point which is, this measure is a retooled measure, so it's based on NQF 1550. One of the things -- so it is at the provider group level. But

when you -- when you compare the scores in the hospital compare measure, both healthcare systems rank no different than the mean. And so there isn't a lot of variation in -- you know, between these two health systems in that measure either.

We -- because we're using -- you know, it's harmonized with that measure, we do believe that if it was used nationally that we would see variation. They see like a nine-fold difference in the lowest and highest-performing sites. But let me just turn it over to Stuart Lipsitz who is our statistician and can talk a little bit about the ICC and some of the other statistical questions that you had. Stu, are you there?

(No audible response.)

Dr. Dykes: Stu, we can't hear you. You're on mute.

DR. LIPSITZ: Can you hear me now?

Dr. Dykes: Yes, now we can hear you. Thank you.

DR. LIPSITZ: I've -- yes, sorry. Yes, I think the -- one issue -- it's not really an issue. The way we split the sample up is we took the -- you know, split exactly the way the random sample in -- in the test and -- and validation group and -- and took the predictive values from the test and used them in the validation group. So I think the high correlations, the way we kind of split the sample up and -- and why the -- the rank correlation is so high between the groups -- so I don't think there's an error in the calculations. It's just that we took a random sample and it just happened that they were highly correlated. But usually the IPC -- I don't think -- you tend to get small ICCs and in this type of study, less than 1 percent -- I think the issue is that, you know, it's not significant means that, you know, it covers zero so that, you know, we just don't have that much variability in the sites in this -- in our study. So -- in the clinician group, and at least in the -- in our sites, it's pretty similar to the hospital

themselves. So because of the way the clinicians work together. So it almost is like the hospital level measure. But yes, I think it's -- I mean, the low ICC is probably the important thing to look at. Or, not the low ICC, the non-significant ICC issues. But as -- as Patty said, I think that, you know, it does -- we did correlate some pretty -- pretty heavily with the Yale group.

Dr. Dykes: The other issue regarding only two EHRs, Cerner and Epic -- that's true that those are the most, you know, commonly used EHRs. That's why we chose them and -- but, you know, the codes that are used to calculate this measure are standard billing codes that would be available in any EHR system because, you know, the -- the providers groups want to get paid. So I agree that it does need to be tested using a much larger sample, and we need much more data, but I am confident that we would find those data in other EHR systems as well.

Ms. Elliott: Thank you so much for your comments. At this point I'd like to open it up for the larger SMP group. So Larry, would you like to address your question or comment?

Member Glance: Sure. The reason I raised my hand earlier was because it was more of a general comment and not -- not completely relevant to this particular measure. But I think to the point that David made about the ICC and how you -- or the fact that you have two completely different values depending on whether you're not -- you're looking at split-sample reliability testing, or you're looking at the ICC it's generated when you're estimating a hierarchical model. They're called the same thing, but -- but they're completely different. And I think -- yes, I think Patrick sort of explained the difference earlier. When you're looking at split-sample reliability testing, it's kind of like a correlation in -- that's what it's called inter-class correlation, but it's -- you're basically comparing the point estimates for

the providers, or for the entities, across two different samples. And when you're looking at the ICC based on the hierarchical model, you're looking at the proportion of the variability that's attributable to the provider. So between providers as opposed to the total variability.

And I think they're two completely different statistical measures. And they will absolutely yield very, very different measures. And they're not at all comparable and you can't average them together. And I think for many of us who have looked at ICCs for surgical outcomes, it really is not uncommon to see very, very low ICCs on the order of 0.03, 0.04 -- maybe a size 0.1. But typically fairly low at the provider level. They're a little higher maybe at the hospital level. But they're not at all comparable. You can't use the criteria that we've used before for reliability. When you expect it to be 0.5 or 0.7. They're just two completely different measures.

And I think Sean and Brian might be able to speak a little bit to the types of ICCs that the folks that do get for the STS cardiac surgery models and what they -- when they're estimating their hierarchical models. Thank you.

Ms. Elliott: Thank you, Larry. We have two other hands raised. Patrick, you are next and then Sherrie.

Member Romano: Okay, I am interested in Sean's response on that question as well. But I have a very specific question for the developers. So in your response you say that an ICC was calculated to describe how much variation in the provider-level scores is due to provider-level signal variation in the 2019 sample resulting in an ICC of 0.726. And then there was some comment in the chat about an ICC of greater than 0.5.

So again there's, you know, apples and oranges going on here. Lots of different ways of estimating the ICC. Can you explain to us exactly what is this

ICC of 0.726, or greater than 0.5? What -- what are the numerator and denominator of that ICC?

DR. LIPSITZ: Yes, actually that was suggest -- I think we sent our -- you know, our patient level of ICC of 0.006 and what are the real suggestions you could, you know -- you get a provider-level ICC, which is a function of the sample size times the ICC of the patient level. And I think the -- you know, different ways to look at it -- but I think the key is that, you know -- that whether or not that patient-level ICC is significant. I mean, they're both functions of the ICC of the patient level. One is multiples by N. So the key is whether that -- you know, the concept if it crossed zero if the patient level is going to cross zero with the formula -- for the provider level, too. So I think one is just a -- you know, a sample. The average number of patients per provider times the ICC. I think it's my N time ICC over 1 plus N times the -- or some function like that. But it is almost like, average number of patients per provider times the ICC. So that's -- so really measuring the same thing. It's one -- so you kind of get an aggregate number at a provider level.

The only other thing I would tell you is also there was just a question about we used ridge -- ridge regression. I know it's not particularly on this, but -- we used it more for the -- I think there were like 30 comorbidities which you knew were highly correlated. So we just did the ridge on comorbidities and not on all -- all predictors. Just the comorbidities because we knew the 30-or-so comorbidities would be highly correlated. So -- yes, thank you.

Ms. Elliott: Great. Sherrie? Did you have a comment or a question?

Member Kaplan: Yes, the -- just as response to Dr. Lipsitz -- the inflation factor is the ICC times the n-1. So it's -- that's different from that -- that's

different from what I was talking about which is the between- over -- the between- versus within-practice variability.

DR. LIPSITZ: Yes.

Member Kaplan: And so it's that -- if you're trying to discriminate between accountable entities, it's the between- versus within-practice variability when you've got a bit denominator in the coefficient of air, then you're going to get a lower ICC. So, you know that's -- that's kind of problematic, especially when you've got things that are confounded by site, like osteoarthritis. So --

DR. LIPSITZ: Yes.

Member Kaplan: -- I mean, I think there are a lot of reasons for this. The question is, when we're dealing with what we're dealing with, what do we do? And we're staring at those ourselves. And I -- do you have collinearity data? Did you run the --

DR. LIPSITZ: No, we -- no, it was mainly -- because we -- we knew the 30 -- we just did the ridge regression penalty on the 30 comorbidities because we knew they were going to -- going to be highly correlated. So we did -- for AIDS and other ones, we didn't really see much collinear -- we don't have the collinearity diagnostics, but we knew a priori that we were going to need to do something. We didn't want to just use a Charlson score. So it was more -- since they're doing a Charlson score, trying to get more information out of the comorbidities to -- to increase the predictability of the model. But, yes -- so I mean, we agreed collinearity at least with the -- an issue with those comorbidities.

But -- but you're -- I mean, that was a good find that, you know -- that the -- you know, the osteoarthritis was different -- involved different sites. Or different -- Cerner versus -- versus ours. So this is important. It means that the -- you know, as you say, you know, the patient population is

different across those two systems. So I mean -- we -- obviously we would love to get more -- you know, the key is to get more systems that are -- or, you know, variety and complication rates to be able to really look at this the -- you know, the measure itself.

Ms. Elliott: Okay. And the last hand-raise that I see is Sean O'Brien. Sean, did you want to comment or ask a question?

Member O'Brien: Yes, I -- well, I comment I'd make other SMP members who are kind of grappling with the interpretation of the statistics. I would say there's more of a connection between the split sample results and the ICC than may be apparent on the surface. The ICC is estimating the correlation coefficient between two measurements of the same underlying true value. It also turns out under some assumptions -- this is also the square of the correlation coefficient between a single measurement and a true value and -- which is also equivalent to the signal-to-noise ratio that we are -- are describing as a proportion of the signal variation.

The terminology is really confusing so I -- you know, what I just said may be completely wrong according to other people. Additional complicating factor here -- the ICC can be describing the correlation between two -- two -- two patients randomly sampled, or a sample average aggregated at the entity level, which, if you have a very limited signal variation, once you aggregate it -- integrate it across enough patients within the entities, you can compensate for that low ICC and ultimately come up with a number that's in the 0.5, 0.6, 0.7 range. So I don't think any number that's -- has an interpretation where a 0.01 is acceptable is very useful for us to be able to make a decision. We need to know, once you factor in the amount of signal variation plus the amount of error you can expect with the sample sizes, then what is that kind of

proportion signal variation going to look like? And I am not sure if I heard that that type of entity level relied over the statistic was provided. So I've missed it. I was thinking we were looking at extremely small numbers. Regardless, we don't have enough information to really know if there's any reliability at either level because of the wide confidence interval. So we're either in a situation where we just don't have enough information, or depending on the answer to some of these questions, we do have a little bit of information. And all the information -- is suggesting -- is very poor reliability. And I -- sorry -- keep going on a little bit more just for -- for Dr. Lipsitz. I wasn't still quite sure how to reconcile a -- a correlation coefficient -- a ranked correlation coefficient of 0.97 with very wide confidence intervals that are in Table 8. It just seemed like -- you know, I may be very much missing something and maybe it's not necessary to straighten out. But it was a remarkable -- a remarkably high degree of correlation in light of all that the -- the ways that the confidence interval suggests.

Ms. Elliott: Thank you, Sean. At this point I do not see any other hands raised. I just want to call attention to the fact that we do have a robust chat that we'll be capturing. But there was a question from a committee member about the sample size of patients, physicians, and hospitals. And Mica Bowen from Brigham Women's provided a response. And there was also some discussion regarding rate of complications and very low surgery. So we'll be sure to capture all of these comments. And there was a couple responses from the measure developer in the chat as well.

So at this point, in light of the time, David I'd like to move it forward for a vote. Do you concur?

Co-Chair Nerenz: I'm sorry, yes. I was muted.

Ms. Elliott: Okay. So with that, Hannah, can I hand

things over to you for the next vote? Or the vote on this measure. Thank you.

Ms. Ingber: Yes, for sure. So voting is now open on Measure 3649-E for Reliability. Your options are high, moderate, low, or insufficient.

(Pause.)

Ms. Ingber: And we're using the same voting link as earlier in the day and we're just looking for subgroup members from Subgroup 1 to vote on this.

(Pause.)

Member O'Brien: I apologize. That link is now in my email somewhere, but I'm having to scroll through to look for it. It may be faster if you're able to resend it, or paste it into the chat.

Ms. Ingber: I can send it just to you, Sean. Yes.

Member O'Brien: Okay, great thank you.

(Pause.)

Member Romano: Having difficulty, sorry.

(Pause.)

Ms. Ingber: We're waiting on just one more.

(Pause.)

Ms. Ingber: Okay, we did need a quorum of 9 voters, and we have a quorum of 10. So I am going to lock the poll and show the results. So voting is now closed on measure 3649-E for reliability. We have zero votes for high, two votes for moderate, seven votes for low, and one vote for insufficient. Therefore the measure does not pass on reliability. Thank you everyone.

Ms. Elliott: Thank you, Hannah. Next slide, please. Either Hannah or Gabby. Great, thank you.

Member Romano: Could I -- I'm sorry. So are we skipping, then, the discussion of validity since the measure did not pass on reliability?

Ms. Elliott: Hannah, can you speak to the next step? Is that the process to move forward?

(No audible response.)

Ms. Elliott: Could we go back one slide? Elisa, could you -- do we need to move on to validity?

Ms. Munthali: We do not need to move on on validity if there were no concerns raised prior to this meeting. And so Hannah, can you confirm whether or not members of the SMP raised any concerns about validity?

Ms. Ingber: They did.

Ms. Munthali: Okay. So you should discuss validity as well.

Ms. Elliott: Okay. So we'll move on to that. So Patrick, do you want to raise -- start the discussion then?

Member O'Brien: Sure. I think that Sherrie really set up the discussion initially with the concerns that she described. But basically, the developers here rely on data element validity. And data element validity in this case was established through a very confusing five-round process. They -- NQF advises developers to apply a gold standard when they're testing criterion validity for EHRs. And so record metrics such as sensitivity and predictive value.

In this case the developers found that they didn't really have a gold standard. So instead they reported measures of agreement -- kappa statistic for example -- through this five-round process. Now that in itself might be understandable. But the five-round process -- it's very difficult to understand because NQF asked developers to report very specifically on the validity of key data elements,

including the numerator and the denominator. And so focusing on the numerator we get through a series of steps, and then we wind up at round five in which there -- a total of 30 patients from one site and 25 patients from another site. And that's the summary of the data element validity that we're provided. Basically a total sample, as I understand it, of 55 from those two samples.

So I found that to be inadequate to assess the data element validity for key data elements. I do appreciate the developers went through a very explicit process of describing all the changes that they made along the way and how they modified each round based on what they have learned in the previous round. But at the end, the sample for data element validity testing was very tiny.

And the second issue really is about the validity of the model. And I think Sherrie has already highlighted this. But there are very severe issues here related to how the model was constructed and the process by which risk factors were selected for the model. And I think the best example is the osteoarthritis, as we mentioned. 97-percent prevalence in one site. 15-percent prevalence in the other site. Clearly these are procedures that are done overwhelmingly for osteoarthritis. We could argue whether or not osteoarthritis should even be in the model. But when you see this magnitude of variation, it really highlights the fact that there wasn't a lot of thought put into the construction of the -- of the model.

I also see, for example, risk factors like pneumonia which is rather odd and led me to look back at the specifications and to find that there was no POA exclusion or limitation for the risk factors in the model. So it raised the question about whether in fact some of these purported risk factors may have been things that arose on the day of the hospital admission -- that they -- the first day of the encounter based on the human readable spec.

So also looking at the prevalence of the risk factors -- again, the prevalence are markedly low compared with the Yale models with similar outcomes. Also no attention to the fact that, when patients present with complications, after a total hip or total knee surgery, they often go to a different medical group. They often go to a location that's more approximate to their homes. And I think the Yale group has found that might be on the order of 20-percent or so of patients who go elsewhere.

So no comment in the documentation about how they attempted to find false negatives related to complication that presented to physicians outside the orthopaedic surgeon's clinician group. So those were just a summary of the concerns related to validity, and some of them have already been discussed or addressed.

Ms. Elliott: Any other comments? Please feel free to raise your hand or -- the information in the chat.

(Pause.)

Ms. Elliott: Sean, do you mind raising your question that you placed in the chat?

Member O'Brien: Well, it's terrible timing because we just voted, but I was trying to get clarification about the interpretation of the ICC that was provided. And I'm seeing a response from Mica Bowen from DWH that says a second ICC was calculated and was greater than 0.50. I apologize because I -- I didn't see that and I was not factoring that into my -- my discussion responses or questions. So I -- I don't know if other people who voted were also aware of that.

Member Romano: Well I -- I did ask that question, and I think that it was answered -- that it was an inflation of the ICC.

Co-Chair Nerenz: But it was not in the written materials.

(Pause.)

Member O'Brien: Well I mean, I think -- they're both okay in the sense that there isn't evidence in the data that suggests high reliability in light of the wide confidence interval. If an ICC estimate is above 0.50 that might be a benchmark that's -- know, been reasonable for other measures. Although there's been a lot of debate about what -- what an appropriate threshold should be.

(Pause.)

Ms. Elliott: So the measure developer responded in the chat to your question, Sean, saying it was in the response submitted for the meeting.

Member O'Brien: I apologize. I didn't read it.

Ms. Elliott: Okay.

(Pause.)

Ms. Elliott: Any other comments on validity?

(No audible response.)

Ms. Elliott: Okay. David, any questions before we move on?

Co-Chair Nerenz: Well I guess it's just a process question. I'm feeling a little time pressure to move on. So this -- they go to -- it's sort of to Patrick and staff. Validity on the original run-through was a pretty clear pass. Patrick has raised some concerns. Question is, does those concerns rise to the level of where we should re-vote this? Or are those just observations that then could be passed on both to the developer and any future consideration, should a standing committee choose to poll this anyway. And I -- I don't know. I don't have the -- I'm trying to sort out, is there a re-vote involved here or not?

Ms. Munthali: Hello, this is Elisa. David, that's exactly right. I think the question we'd like to pose

to the SMP is if, given the recent discussion, would your vote change? Or would you accept your original vote coming into this meeting -- on validity?

Member Romano: There is a formal request. I'm -- I'm moving to request a vote -- a formal vote on validity.

Co-Chair Nerenz: With that in mind, I'm inclined to do that. I mean, obviously Patrick's concerns are -- are sort of eloquent and well-stated. It doesn't take us too long to re-vote. And I -- I feel like it's more appropriate to do that than to simply ignore what Patrick just said. And certainly others can weigh-in on that, but I -- I don't know, I don't feel like on my own discretion I can say yes, no. But I am included to say, if there's a formal request that we re-vote -- the lean would be to do it.

Ms. Elliott: If the committee agrees, we do have the information available. Before we -- we do that, I'd like to give the developer a an opportunity to respond to the validity.

Co-Chair Nerenz: Yes, indeed. Yes.

Dr. Dykes: But first, in terms of the gold standard, the issue that we found was that there really isn't a gold standard because we found with the eCQM, with the electronic health record, sometimes the same code is used if a person is being followed because they have a history of something as if when they actually have the condition.

And so there's some error there. And so that's not a gold standard. And usually, you would say, well, the provider or the clinician doing the chart review is the gold standard, but we also found that sometimes there is information that's buried within the chart that clinicians will routinely miss.

And so we found errors in both and didn't feel that either was a clear gold standard. And so that is why we did the cap-all. We did do -- even though the

last round was 55, we did hundreds of chart reviews. I mean, if you look at -- and they were all random samples, and then each random sample, we included people from the numerator, the denominator, and people that were excluded.

So I do think that we pay very close attention to our validity testing. I wonder if, Stu, you want to add anything to that before I talk about the risk adjustment model.

DR. LIPSITZ: No. I think that's right. I mean, I think the sample size is 30 in the numerator, and I forget -- 30 or whatever for the denominator. I mean, based on how high the capital was, at least it seemed like we're pretty good at getting really good agreement for that.

Dr. Dykes: And then, with regard to the risk adjustment model, we harmonized with NQF 1550, so the osteoarthritis and the complications, that's done exactly how the Yale CORE group does a risk adjustment model.

The modifications that we made in the risk adjustment model were based on a review of the literature that showed disparities for certain groups with these measures. And so we ended up adding some of the socioeconomic, social determinants of health, or proxy score, to the risk adjustment model. And that was based on the literature.

So there was a solid foundation for what we included. We also, through advisement of our TEP, added BMI because they thought that that was important to add. So that is my response about the risk adjustment model.

DR. LIPSITZ: I agree with Patty. I think the elements we used in our prediction model, we tried to, with the TEP, internally pick the variables he thought were most predictive. I mean, like you said, instead of using the Charlson score, we also thought the individual comorbidities and some other way of

combining those, like in the ridge regression approach, was the best way to go.

I mean, I think the issue -- well, you know, the important issue about this, the differences between the comorbidities in Cerner versus our site, which is nothing -- it's good to have that diversity, but are those -- it would be nice to have more sites with more diversity and those kind of characteristics.

Member Romano: Could you briefly address the question of whether you were able to count only conditions that were actually acquired after the surgery? Because pneumonia seems extremely unlikely to be present before the surgery -- pneumothorax also extremely unlikely to be present before surgery.

Dr. Dykes: Yes. We did include only conditions that happened after the index event, or the procedure.

Member Romano: I'm talking about the risk adjustment model.

DR. LIPSITZ: So you're talking about what we -- I think it had to be before the admission, right, Patty? I mean, I think he's talking about when pneumonia occurred. We didn't count it if it occurred during surgery or after. It had to be before surgery, I think. Other comorbidities, how we -- there was the previous 30 days or previous --

Dr. Dykes: Well, the comorbidities, yes, in the previous 30 days. They had to occur before the procedure date.

DR. LIPSITZ: Yeah.

Member Romano: It seems like you allowed for them to be present on the procedure date. You might want to double-check that.

DR. LIPSITZ: Yeah, I think we have to check it. I'm pretty sure it had to be prior to the procedure date, but we'll double-check.

Ms. Elliott: Sherrie, you had your hand raised?

Member Kaplan: There's just one more issue. There seems to be a tension between trying to harmonize this measure with 1550 and trying to -- and you can't adjust a confounder. So when you get that kind of level of confounding, it really does cause you trouble.

So, statistically, there is kind of a tension that you're struggling with, I think, that may have played a role in how this played out.

DR. LIPSITZ: Well, yeah. I mean, I think the issue is, if it's so different across sites, there's no way to really adjust for it. I mean, whether you try to adjust for it or not, if they're so different across sites --

Ms. Elliott: Right. That might --

(Simultaneous speaking.)

DR. LIPSITZ: -- that's what the difference is. Yeah, that's causing difference.

Member Kaplan: That might -- back to Patrick's point about not including it.

Member Romano: Yeah. I mean, it's really an issue that the data are invalid from the Cerner site. I mean, there's no way you could have any population of patients having total hip and knee surgery with only 15 percent osteoarthritis. It's impossible. So it's a data validity issue from the Cerner site.

DR. LIPSITZ: Okay. That makes sense. It's more like a measurement error problem, so not including it. If it's all a measurement error, then we wouldn't want to include it. Yeah.

Ms. Elliott: Okay. There was a question in the chat.

Larry, did you want to raise your question?

Member Glance: Sure. And my question was does the measure exclude patients with fractures? And the reason I'm bringing that up is fracture patients certainly could present with pneumonia.

And if it does exclude patients with fractures, then you're probably -- as Patrick was saying, there's got to be some kind of data quality issue here because you're not going to operate on people who present with a pneumonia or pneumothorax for an elective hip or knee replacement. That's a problem with the data-cleaning piece.

Dr. Dykes: It only includes primary total hip/total knee, which means that they would not have had a fracture.

DR. LIPSITZ: I mean, the real point at issue is, I mean, we could really do our checks for the Brigham -- for the MGB data. We didn't have quite as much flexibility over the Cerner data, which I guess it sounds like people are worried about.

Member Romano: Ballpark, you say the total hips/total knees are done for osteoarthritis or for rheumatoid arthritis in general, and so the sum of those two should equal about 100 percent. So you have 97 percent OA. You have about one percent RA. That sounds low.

But at the Cerner sites, you have 15 percent and 0.7 percent. So you have unexplained 85 percent of all the total hips and knees.

Dr. Dykes: Well, we did ask Cerner about this, and they said that this site had recently changed to a different system, and they knew that there was a problem in that coding.

Ms. Elliott: Okay. I don't see any more hands raised, and no new questions have come into the chat.

So, Patrick, I defer back to you, as you're the one

that raised wanting to revote on validity. Does that still stand?

Member Romano: Yes. I mean, I think it would be a useful signal.

Ms. Elliott: Okay.

David, can I turn to you to recommend voting?

Co-Chair Nerenz: Yeah. Please, let's do it. Relative to time spent discussing, this won't take us much more. We should do it.

Ms. Elliott: Okay.

So, Hannah, I think I'm handing things off to you to initiate the voting on this measure for validity.

Ms. Ingber: Yes. Thank you, everyone.

Voting is now open for Measure 3649e on validity. Your options are high, moderate, low, or insufficient.

Looking for just one more.

(Pause.)

Ms. Ingber: All right. We have all our responses in, so I will lock the vote and show the results.

So voting is now closed on Measure 3649e for validity. We have zero votes for high, six votes for moderate, four votes for low, and zero votes for insufficient, for a total of ten votes. Therefore, consensus is not reached on validity.

Thank you, everyone.

Ms. Elliott: Thank you, Hannah.

We'll now move forward to Measure Number 3650e. And I'll be handing things over to LeeAnn White, an NQF Director.

Team, do we need to check quorum or attendance at this point?

(Pause.)

Ms. Elliott: Hannah, do we need to pause to check that, or are we good to continue?

Ms. Ingber: I think we're good to continue. But if anyone needs to leave, just please chat us so that we can keep mindful of that.

Ms. Elliott: Excellent. Okay.

So, LeeAnn, if you're ready, go ahead on Measure 3650e.

Ms. White: Wonderful. Thank you, Tricia.

Okay. So, moving on for our next surgical measure, we have Measure 3650e, risk-standardized inpatient respiratory depression rate following elective primary total hip arthroplasty or total knee arthroplasty. Measure 3650e is a new measure for the fall 2021 cycle, and the measure developer is Brigham and Women's Hospital.

The eCQM estimates the risk-standardized inpatient respiratory depression rate following elective THA and/or TKA at the clinician group level for adults 18 years and older across all payers.

This outcome measure is analyzed at the clinician group practice level and is intended for the inpatient hospital care study. The type of score is rate proportion, and for risk adjustment, this measure uses a statistical risk model with ten risk factors.

Moving on to the reliability, this measure passed with a moderate rating for reliability. The developers conducted reliability testing at both the patient encounter level and the accountable entity level. For patient-encounter-level testing, the developers used 30 random patients to evaluate the accuracy of the eCQM extraction. The developers

also compared the sociodemographic characteristics of patients included in the test and validation samples and found no difference between sites or clinician groups.

Some SMP members raised questions on whether this reliability testing method of sociodemographic characteristics across two subgroups demonstrates reliability. At the accountable entity level, the developer performed reliability testing at the measure's core level using a test-retest approach to examine the reliability of the predicted expected ratios at the clinician group level.

The developers found that the test and validation samples gave a similar ranking of the 17 clinician groups with respect to the predicted ratios, with a Spearman's rank correlation of .767 between the two samples.

The developer also estimated that the intraclass correlation between the clinician groups and the ICC value at .069151. Some SMP members noted the high correlation statistics that raise reliability concerns with a low ICC value.

For validity, validity was tested at the patient encounter level and the accountable entity level and received a low rating from the assigned subgroup. So this measure did not pass validity.

For patient encounter validity testing, the developer assessed the frequency of data elements needed for risk adjustment and data element agreement between manual chart review and EHR calculation.

At the accountable entity level, the developer convened a TEP to assess the face validity, reporting that three out of seven, 42.86 percent, of the TEP members agreed that the measure was actionable to improve quality of care.

The developers also risk-adjusted the predicted and expected numerator events for age, gender, type of

surgery, insurance, rates, household income, English as a primary language, smoking status, BMI, and comorbidities.

The SMP members raised concern with the validity testing, primarily with the low results of the face validity testing. Several SMP members raised concerns about the conceptual rationale for the risk adjustment strategies, and SMP members noted that the use of social risk factors not used in this measure without a robust conceptual frame for why these -- or what the proxy might influence in patient respiratory depression.

So, with that, I'm going to hand it over to Dave Nerenz to open discussion around some of the concerns related to both reliability and validity.

Co-Chair Nerenz: All right. Thanks. We're obviously quite a bit behind time, so let me just turn very quickly to staff on this. Let's focus exactly on why this was posed for discussion.

We do not have a CNR issue to resolve. There was a clear pass, or at least it made a pass, on reliability. There's no pass on validity. What precisely should we be talking about here? And then we can lead the discussion that way.

Ms. Elliott: Since the validity was no-pass, that would be the priority. But it was also pulled for discussion for reliability, but we'll start with validity.

Co-Chair Nerenz: Yeah, that's fine, because I know occasionally people express concerns, but it doesn't rise to the level of pull for discussion. All right. So let me focus directly, then, on validity, and I'll try to cut to the chase here.

As you can see on the screen, the votes were pretty strongly in the area of low. And I can speak to that a little bit. The measure developers did two levels of validity testing. They did patient-level validity testing, focusing on data elements. They did entity-

level validity testing that came from votes of a technical expert panel.

And I'm speaking for myself now. There are concerns at both levels. As we just heard, the technical expert panel, actually, a minority of those voting said this was the national gold quality measure. I don't think we've ever seen that before in a measure submission. Since the entity-level validity rides on its face validity, that to me was a very strong concern.

And then the other point's a little more subtle. For the data element validity, the issues were largely about agreement between the information in the electronic medical record and then in a sample of records that were hand-extracted.

And the developers pointed out that the concept of gold standard didn't feel comfortable to them because there were potential errors on both sides. And I'll let my colleagues speak to this. In my own sense, if you don't have a gold standard, then you're really talking about reliability of the data elements, not validity.

And so, in my judgment, then, there wasn't strong evidence of validity of the data elements, nor was there strong evidence of validity of the measure score. So why don't I pause there? And let's see what other members of the subgroup have to say.

Member Romano: This is Patrick. I'll just point out that the developers did explain in their response that the reason that their scoring was so poor by their technical expert panel was that the panel believed that the accountable entity essentially should be the facility rather than the clinician group because it's the facility that provides all of the support services, the nursing services, the respiratory therapy services, so forth, that would be linked to reducing the respiratory depression rate.

So that seems like a plausible concern. But, of

course, we are asked to evaluate the measure as it's presented to us, which is a clinician-group-level measure. So the technical expert panel voted against the measure, and I see no reason to second-guess their vote.

(Pause.)

Ms. Elliott: There are no other hands raised. Larry has a question in the chat, but we'll address that when we open it up to the full committee.

Any other SMP subgroup members care to comment?

(Pause.)

Ms. Elliott: We'll move to giving the developers an opportunity for a response.

Dr. Dykes: Right. So we explained the gold standard issue with the last measure. It's the same with this one. And we agree that the TEP did not vote in favor overall with this measure because they were worried about the level of attribution. They thought it would be a good facility measure but were worried about attribution at the provider group level.

DR. LIPSITZ: The only other thing I'll add is about the second bulletpoint about the reliability. I mean, we didn't put comps on the Spearman correlation coefficient, which would be pretty wide because I think you only have 17 sites.

So you have both comps on the ICC, even though it was discussed in the last discussion. They are measuring different things, the Spearman for the split sample. But we didn't put a comp on that, which would be pretty wide with the low number of sites we had.

But I think the Spearman correlation was still about .75, which is still pretty good for this measure. But still, we didn't have a comp, so it could be a five -- you know, 90-something percent, as low as 50. So

have to take that into account to.

Ms. Elliott: Thank you.

I do not see any other hands raised.

Larry, would you like to raise your question as part of the SMP?

Member Glance: Sure. I was just wondering -- and this is a question for the measure developers -- about how inpatient respiratory depression was defined.

It's a bit of an uncommon respiratory outcome to look at. I mean, people might look at postoperative pneumonia or reintubation rates or readmission to an ICU, but I'm not sure how you capture respiratory depression even using the EMR.

I mean, I suppose you could do it, but it's just not a very good outcome to follow. And I would imagine that that might have been, also, one reason why the TEP might have been not in favor of this particular measure.

Dr. Dykes: So we used documented diagnostic codes related to respiratory depression or procedure codes from mechanical ventilation, intubation, or repeated oxygen saturation readings that were less than or equal to 88 percent and greater than 30 percent within a period of 24 hours of the procedure or during the inpatient stay.

DR. LIPSITZ: And also, we had a lot of discussion about exactly the question about how to measure this. And through our discussions, what we narrowed it down to is what we felt is as valid as possible way to measure.

Dr. Dykes: TEP agreed with the definition. They just felt --

DR. LIPSITZ: Yeah.

Dr. Dykes: -- the attribution to be the facility level.

Member Glance: That makes a lot of sense. Thank you.

Co-Chair Nerenz: Tricia, just to move us along, are there any other hands raised that you see?

Ms. Elliott: Yeah, just one more. Zhenqiu just raised his hand.

Member Lin: Yeah. I just have a question for the developer.

So the TEP's concern is that this should be a facility measure. So did you ask them whether surgeon -- a group bear responsibility deciding where to take the patient to have surgery? I imagine most of the patients, when they go to surgery, each patient choose where to go to.

Dr. Dykes: We had that discussion specifically. So in the context of our healthcare system and also the healthcare system geographically distant, the surgical groups were affiliated with the hospital where they did the surgeries. So maybe it was more clear cut.

Member Lin: Okay.

Co-Chair Nerenz: And, Tricia, I'm risking overstepping my bounds here, but we're way behind time. And as I look at the three bullets on the bottom, it seems to me that we talked about the testing sample size in the last measure. I don't think that's a hot issue.

The developer response and reliability, I think, strengthened what they had originally, but that was a pass in the first place. I haven't heard anything that I think would change anybody's view on the validity. I'm inclined to say let's move on. I'm also certainly willing to listen to other objections, but we've got to keep going on our agenda.

Ms. Elliott: Okay. I just want to call out in the chat the measure developer addressed a question about the respiratory depression. So --

Co-Chair Nerenz: I mean, it's fine, but it's not an issue in front of us, and it's not something that there's -- this is being pulled for revoting on.

Ms. Elliott: Okay.

Co-Chair Nerenz: It's important to know, but I think we're done.

Ms. Elliott: Okay. So move to vote on validity, then, or just move on?

Co-Chair Nerenz: Move on. I would not revote it. Again, others can object, but I haven't heard anything that fundamentally changed our assessment.

Ms. Elliott: Okay. Very good. I'm not hearing any objection, so we'll continue to move forward with the next measure.

And this is Measure 3652e. LeeAnn White is the NQF Director, so I'll hand things over to LeeAnn for this measure.

Ms. White: Okay. Thank you, Tricia.

So this measure passed both reliability and validity but was requested to be pulled for discussion. So this measure is the risk-standardized prolonged opioid prescribing rate following elective primary total hip arthroplasty and total knee arthroplasty.

The measure passed, again, both reliability and validity with moderate ratings. This is a new measure for fall 2021, and the measure developer is Brigham and Women's Hospital.

This eCQM assessed percentage of patients 18 and older across all payers who were not previously exposed to opioids within 90 days before the

THA/TKA procedure and who were prescribed opioids for greater than 42 days following an elective primary THA/TKA.

This process measure is analyzed at the clinician-group-practice level and is intended for ambulatory care, inpatient hospital, and outpatient service care settings. And for risk adjustment, this measure uses a statistical risk model with eight risk factors.

So the developers tested reliability at both the patient encounter and the accountable entity levels and received a moderate rating. At patient encounter level, the developer used the eCQM feasibility scorecard to assess the EHR data availability, accuracy, terminology standards, and workflow. And the scorecard received a measure score of 1 out of 1 for all 22 data elements.

The developer also compared sociodemographic characteristics of patients included in test and validation samples and found no difference at the patient level or between clinician groups. Some SMP members questioned whether comparing these factors across tests and validation samples adequately demonstrated reliability.

For the accountable-entity-level reliability testing, the developer used the test-retest approach to test the reliability of the predicted expected ratios at the clinician group level. The developer estimated how the two random samples agreed using the Spearman correlation coefficient, which was found to have a value of .8182 for THA and .8909 for TKA. Lastly, the intraclass correlation for THA was .0929, and the ICC value for TKA was .11675.

Regarding accountability-level testing, SMP raised concerns with the ICC results and whether the measure can capture variation and provider performance.

So, moving on to validity, this measure received

moderate ratings for during the initial evaluation. For the patient encounter level, the developer analyzed extracted patients' EHR data, compared the findings to the results turned in by the eCQM, and found that the manual chart review and the eCQM had a perfect agreement, a CAP of one.

Validity at the accountable entity level was tested through face validity. There was a seven-member expert panel with 100 percent agreement that the performance scores resulting from this measure can be used to distinguish good from poor clinician-group-level quality related to patient safety.

And lastly, the developer adjusted, predicted, and expected extended-use rates for age, sex, race, household income, English as a primary language, BMI, and comorbidities. The same statistic was used to assess the model strength and predict the prolonged prescribing events, and it was .708 for THA and .655 for TKA.

Before I hand it over to Jenn Perloff for further discussion on validity and reliability testing, there were also a few concerns raised by the SMP during the initial review of this measure. One member raised a concern with the 42-day interval selected for the measurement and whether the time starts following the surgical procedure date or on the date of discharge.

Some SMP members raised concern about the EHR and whether this measure could be generalizable to EHRs outside Epic and Cerner. One SMP member noted that no data elements are missing in Epic, but days supplied is missing in about 34 percent of the time in Cerner.

SMP members were also concerned about how this could be generalized to other EMRs. And then, lastly, several SMP members questioned whether the measure is appropriately categorized as an outcome measure rather than a process measure.

So, Jenn, I will hand it over to you to discuss the concerns related to reliability and validity.

Member Perloff: Great. I will try to be quick, and I'm going to be more on the reliability side.

But the first thing I want to say, excess opioid prescribing is obviously a very important issue. So I want to highlight the importance of this measure, and anyway, enjoyed reading it.

One exclusion I want to raise, we kind of reflexively accept the exclusion of discharge against medical advice, AMA. I would challenge us all to think about whether those cases should actually be excluded. I understand their missing data concerns, but we really have to ask ourselves, why did those patients leave AMA? So I just wanted to raise that.

I raised the issue I had around reliability, generalizability, with using the two high-end EHRs of the world when there are many more dispersed, less sophisticated EHRs that many hospitals and delivery systems use.

I was concerned about the liability testing, about the pulling across four years for the analysis. When I think about EHR data, I'd be interested in reliability year over year in addition to within and across providers and data elements and all the other different dimensions. So not considering time was a concern for me.

And then, obviously, we heard throughout the discussion of this measure set the issues around the ICC, and that was one that came up for me as well since those were particularly low, although I've been educated in this discussion today about differences in surgical measures and surgical groups compared to some of the other things we look at.

So I would just throw those initial issues out, but definitely would turn to my colleagues to enhance. And I'm not sure who asked the measure to be

pulled, but I think that would be an important person to hear from.

Member Needleman: I wasn't the first to ask it to be pulled, but I certainly endorsed that. And it was over the risk adjustment model.

Member Perloff: Great.

Member Needleman: And very simply, yeah, this is a process measure. And we've had lots of discussions in the field, in the Assistant Secretary for Planning and Evaluation's Office, in the literature, about when it is appropriate to address for sociodemographic factors and when it is not.

And the risk adjustment model here has race and ethnicity. It has an area-level income measure. And I see no justification in the documentation for why those are legitimate factors that should influence opioid prescribing.

Therefore, I just don't think that the -- this is a case where the inclusion of sociodemographic factors makes it a fundamentally flawed risk adjustment model.

Member Perloff: Excellent point.

Member Romano: And I'll second Dr. Needleman that this is really foundational. This is a process measure that is intended to capture when physicians make incorrect decisions related to the process of care, as in prescribing opioid medications for an excess period of time.

Now, we can argue about whether such measures should be risk-adjusted at all. The vast majority of process measures are not risk-adjusted at all because we define the cohort of interest using exclusions or using stratification so that we say, for this category of patients, it is always appropriate to do this, or it is never appropriate to do this.

So the vast majority of process measures are not

risk-adjusted. There may be some circumstances -- for example, if patients have other chronic conditions that cause chronic pain, one could argue that that should be included in the risk-adjustment model. We found, for example, in this case that cancer, although I thought it was an exclusion, it was actually included in the risk model.

And that's something that could be justified. But certainly, social factors cannot be justified. It's just fundamentally wrong to do this in a process measure. And the literature that the developers cite is all related to outcomes where we could have some debate about whether social factors should be in an outcome model or not, but not in a process model.

And I'll just highlight one other point, which is about the missing data. A very important problem with up to 35 percent missing data from the Cerner sites -- why is it missing? It's missing because people aren't prescribing narcotics through the same electronic health record system that's being used to harvest the data.

So if you give a paper prescription, for example, it doesn't appear, and there's no ability to link that to the electronic health record. So this gets to Dr. Perloff's point about generalizability, that in a major academic center with a fully integrated system that's using a state-of-the-art Epic system for electronic prescribing, we have no missing data or virtually no missing data.

But in a real-world setting where people are writing prescriptions on paper, where people are writing prescriptions by telephone or outside of the electronic health record, we have a lot of missing data. And it's problematic to assume that these people didn't get opioids, which was the assumption made here, the imputed value of zero.

So those two, I think, are the issues at hand: the exclusion and the risk adjustment.

Ms. Elliott: Any other subgroup/committee comments or questions?

Member Perloff: Just a technical question about this issue. I think this was labeled an outcome measure, but there's this consensus building that this is a process measure. Is that also something that we can provide feedback to the developers on? It just seems like if it's miscategorized, that's a key feature here.

Co-Chair Nerenz: Yeah. If I read correctly, in the developer response, it seems they state clearly themselves it's a process measure, unless I read that incorrectly. I don't think it was submitted that way originally, but I think I see that in the developer response.

Member Kaplan: This is Sherrie Kaplan. I was also confused about the issue of use of a prescription. So if it's a prescription for longer than 42 days, that's clearly a process measure. If it's used longer than 42 days, then that's outcome -- it sounds to me more like an outcome measure.

Member Romano: And I'll just point out that this also -- this ties to the issue -- this is really -- almost all of our prescribing measures that we've reviewed are measures that are based on pharmacy claims or based on patient-reported data.

So this is unusual to rely on the prescribing information coming from the physician's record in the electronic health record. And it just raises a whole other option of gaming, right, where if you wanted to lower your rate on this measure, all you would have to do would be to prescribe by telephone or prescribe on paper and not document it in the EHR. So that's more of a usability concern, but it's tied with this validity issue of missing data.

Ms. Elliott: Sherrie, you had your hand raised?

Member Kaplan: Yeah, just one more follow-up on

Patrick's point about the differences between the two electronic medical record systems. The prescribing system sophistication doesn't account for the fact that smoking status was missing for 97 percent of the Cerner sites.

So that's probably part of the routinely reported information that comes back from the history. So it's not just the prescribing behavior that's a concern here. It's also some of the data that were actually recorded and included in the response.

Ms. Elliott: Okay. Any other subgroup comments before we go to the measure developer?

(Pause.)

Ms. Elliott: Okay. It looks like, Patricia, you and your team on --

Dr. Dykes: Last of my team still standing.

So, yeah, thank you for your comments about the risk adjustment model. We did have a lot of discussions about this with the TEP, and you're correct that the social determinants in the literature are linked to outcomes of these patients.

The TEP agreed that they should also be included in this process measure. However, we did provide in our materials both the adjusted and the risk-adjusted, and you can review the results of both. I think you'll see even more meaningful differences in the process if you look at the unadjusted results.

In terms of the missing data, for opioids, my understanding is that these have to be prescribed electronically. And so the problem that we found is that in one healthcare system, they were using templates where the required data elements were required, and they were prescribed uniformly. In the other Cerner site, they didn't have a requirement that all of the aspects of the prescription had to be recorded. And so that's where we saw missing data.

But see, part of the -- we think that the value of this measure is improving those practices. Like if we are going to increase our accountability with opioids, we have to be able to measure how long we're prescribing them. And unless we get on the same page about what are the data elements that need to be routinely captured in the EHR, then we're not going to get there.

And so these are the kinds of discussions we had with our TEP. And so they thought that despite the missing data, that this was very meaningful because already, after our discussion, when we presented the results to the Cerner site, they said they were going to make changes in their EHR because they hadn't realized that they were missing all of these data and what the implications are.

So, in terms of smoking status, that missing data, the issue at the Cerner site was that they had changed the question that they used for smoking status over the last year or so, and so it wasn't reliable for that reason. So I think it brings up the issue with, if this is an important data element, we all should be collecting it the same way. People should be using a standard data element.

And they had switched to the same data element that we were using in our system, but we weren't able to capture it for a lot of the patients in part of the sample.

Ms. Elliott: Thank you.

Member Romano: I'll just say I put it in chat that it's simply not true that opioids have to be prescribed electronically. Where did you hear that from, or where did that idea come from?

Dr. Dykes: In our health system, if you don't submit it electronically, the pharmacy won't fill it --

Member Romano: Oh, that's a closed --

(Simultaneous speaking.)

Dr. Dykes: -- in Massachusetts. So there's more and more national reporting where they're trying to track, and this is one area where in order to cut down on people going to different states to get opioids, there's tracking systems. And many states have those and require that the opioids have to be administered electronically so that they can be tracked. And to help with compliance, a lot of the pharmacies won't fill it unless it's submitted that way.

Member Kunisch: Hi. This is Joe. I think this is a great discussion. In the state of Texas, it is now required electronically. So I think this discussion brings out a lot of things because I've been involved in multiple eCQM testings, and when a developer comes in and it's testing a new measure out, we're using whatever yields or data is available at the time.

It's completely different when an eCQM is approved and then required for certification, because then the vendors will build out specifically to capture that data. So it makes it somewhat easier. Still challenges, but then you'll actually get reliable data because a vendor will say, here's the three places you can document this which will count towards the measure.

When they're coming in new, it's completely different. I mean, it's dirty data, I call it. And we do a lot of cleaning and scrubbing to get a good data set for a measure developer. So this is always going to be a challenge for eCQM developers going forward.

Member Perloff: I think you made my point about the number of EHRs this is tested in perfectly.

Ms. Elliott: Jack, you have your hand raised?

Member Needleman: I do, and I'm going to preface

this -- it's a question for Patricia, and I'm going to preface this because I think, overall, the Brigham and Women group has done a good job, given the data they had available, in developing the concepts of these measures and trying to implement them.

And I've got problems with them in some cases, but I do want to make that statement because I'm about to be extremely rude, which is, Patricia, what did you think you were measuring about clinical decision-making when you included race and area-level income in the factors in the risk adjustment model that should influence whether or not a long-term, prolonged opioid should be prescribed? Because that, to me, is at the core of our whole discussion about when social determinants are appropriately included and when they're not.

What's the clinical justification for including those in the risk adjustment?

Dr. Dykes: Harmonizing the risk adjustment model with our other measures. And so, after participating in this discussion, I can see that maybe it was the wrong decision, although I have to say that we had many discussions about this with our TEP, and they didn't disagree with it. In fact, they agreed with the risk adjustment model.

But yeah, I can't speak to that. But I can say that we did include the results without the risk adjustment, and those are available as well. I think the measure performs -- it does show differences whether you risk-adjust or not.

Member Needleman: Yeah. Our problem is we have to choose to vote up or down on the measure as it's been spec'd to us.

Co-Chair Nerenz: Yeah. I'm wondering -- I've got sort of a thought here on process, and I'm combining what we just heard the last few minutes about having both unadjusted and adjusted data in front of us. And also, that came up in the developer

response as well.

In fact, there was a developer suggestion that they'd be willing to go with NQF's guidance about whether this went forward either as unadjusted or adjusted. Now, I haven't had a chance to go back and review the original materials in detail, but I think there's been enough raised here that a revote seems justified.

But I'm wondering if what we might need is an overnight as a group to look at what has been given to us on the unadjusted rates as a basis for a revote, just based on what we've just heard from Patrick, from Jack. I share the same concerns.

And what I don't know -- I just can't read it fast enough -- is whether there's enough presented to us that we could vote on the basis of unadjusted reliability and validity statistics and then make it clear that that was the basis. I see that as a possible path forward, but I could be corrected by somebody who already knew that perhaps we can't do that.

I just don't think I'd be prepared to revote right now if what we're really asked to do is think about this as an unadjusted rather than adjusted measure. I'm looking for process help here.

Ms. Elliott: So I would have to ask the developer if there's unadjusted testing that was provided that might add clarification.

Co-Chair Nerenz: I think that's what I just heard, although it'd be nice to hear that confirmed. And that was the basis for what I said.

Dr. Dykes: Yeah, we did. We provided both the adjusted and the unadjusted rates.

Ms. Elliott: Got it. Okay. So then --

Co-Chair Nerenz: And --

Ms. Elliott: Go ahead, Dave.

Co-Chair Nerenz: I'm sorry, Patricia. Let me just press the point. Would we be able to find reliability and validity statistics based on unadjusted rather than adjusted rates, or are those tests only done on the adjusted?

Dr. Dykes: No, we did it on numerator, denominator -- we did look at elements of the risk adjustment model, but we did it for the whole measure.

Co-Chair Nerenz: All right. Well, I'm still a little confused, but I guess I feel very uncertain about just moving directly to a revote, but I also don't like the idea of not revoting. I'm wondering again, as an alternative to either of those two -- we've perhaps taken enough of the developer's time, and we've heard the responses.

If there's going to be a revote, I wonder if we should slide it into our agenda tomorrow when we've had a chance to perhaps look at some things overnight and perhaps form different opinions.

Ms. Elliott: That is definitely an option, so I encourage other SMP members to weigh in.

Member Needleman: I'm prepared to look at stuff tonight and vote tomorrow.

Member Romano: It's certainly an option. I mean, what's interesting is that their risk model for total hip does have a C-statistic of .708. We don't know how much of that is driven by the social factors that we don't like and how much of it is driven by clinical factors which might have some justification.

For example, the same model includes factors like skin ulcers, which might be painful; major psychiatric disorders, which might be associated with an increased need for long-term opioids -- certainly osteoporosis, other bone cartilage disorders. So there are other clinical factors in the

model.

So it's a little hard to embrace this as an unadjusted measure when the developers have made an argument that they need to adjust to account for patient characteristics that might justify longer courses of postoperative opioid therapy.

Co-Chair Nerenz: Good point. Good point.

(Simultaneous speaking.)

Co-Chair Nerenz: -- a pendulum swing or a Goldilocks. You can do too much adjustment, you can do no adjustment, and the right thing might be somewhere in the middle. But we don't have the middle to vote on. So that's tough.

Member Romano: And it's a little puzzling. Maybe the developer could answer, but -- because they say they excluded patients with advanced cancer, but then metastatic cancer is actually in the model as a risk factor.

So I would've thought that those patients would be excluded because, often, those patients require -- it's accepted management that they would receive long-term opioid therapy. So maybe the developer could address that.

Dr. Dykes: Also excluded anyone who received opioids in the last 90 days, so a lot of those people would fall out because they would have had opioids in the last 90 days.

Member Romano: But then why would metastatic cancer be a risk factor in the model?

Dr. Dykes: I have to check on that.

Ms. Elliott: So I think, Dave, if I can turn it back to you, I think I'm hearing that SMP is okay with deferring the vote till tomorrow based on your proposal.

Co-Chair Nerenz: I've heard no objection. Of course, it's hard to interpret that. I'll make that motion; see if it flies.

Member Kaplan: Second.

Ms. Elliott: I don't see any hands raised nor any objections in the chat.

Co-Chair Nerenz: Let's try the --

(Simultaneous speaking.)

Co-Chair Nerenz: We still may have to clarify for ourselves when we pick this up tomorrow. What exactly, then, are we voting on? Are we voting on the adjusted version? Are we voting on the unadjusted version? Or don't -- for some reason, after a night's thinking, we don't want to do either.

But again, I think something we do tomorrow will be better informed and more valuable than what we do in the next few minutes.

Member Romano: Right. Just to -- because we might need some advice from Matt Pickering or Elisa Munthali because we have before us a measure that passed, barely, on reliability and validity. And we've heard a lot of concerns about validity in particular.

And we've heard the idea that a way to address these concerns would be to vote on the unadjusted measure, which was not the measure that was officially presented to us. So in order to do this, how do we proceed in terms of the process?

Ms. Munthali: Hi. This is Elisa. You are right to be confused, because our guidance to committees, including the SMP, is to vote on the measure as specified. What is confusing for you is that it's specified in two different ways, which could render different decisions.

What we're proposing, and I think it's what Dave was articulating to the team and also Tricia, is that

we speak with the developer and summarize what we're hearing, the concern we're hearing from the committee, and see if the developer is able to make any changes and present the measure in a way that follows NQF's process and the application of our criteria.

We want to make sure that tomorrow when you meet, it is an efficient yet thorough discussion. And we really appreciate all of the feedback you've given us so far. But we appreciate the challenge you're going through. We're taking diligent notes, and we're trying to come up with a path forward.

But at this immediate juncture, we believe that we should hold discussion, work with the measure developer, and come back with a proposal so that you have clear guidance on how to proceed.

Member Romano: We're extremely -- thanks.

Ms. Elliott: Okay. So that is how we'll proceed on this particular measure. We did have an afternoon break scheduled.

Dave, I defer to you and Christie if perhaps we want to take a quick five-minute break. And then we have two more measures that we will try to continue discussion on for this afternoon, I believe. Or is it three? No, two -- three measures. And we'll see how far we get and may need to defer some discussion to tomorrow.

So that's my proposal to you and Christie, perhaps.

Co-Chair Nerenz: Yeah. Let's try to do a very brief break. No matter how dedicated we are, there are certain biological imperatives that we have to deal with. Five to ten minutes. Five's tough, but let's see how close we can get.

Ms. Elliott: Okay. So it's 3:29. Why don't we do ten? 3:40 we'll reconvene and see how many measures we can get through. Thank you. So please

reconvene at 3:40 p.m. Eastern Time.

(Whereupon, the above-entitled matter went off the record at 3:30 p.m. and resumed at 3:41 p.m.)

Ms. Elliott: Welcome back, everyone. It's 3:41, so I think we'll jump in and get started.

## 3638 Care Goal Achievement Following a Total Hip Arthroplasty (THA) or Total Knee Arthroplasty (TKA) (BWH)

The next measure up for discussion is Measure No. 3638, and Matt Pickering from the NQF staff will be starting with the background on the measure.

Dr. Pickering: Great. Thanks, Tricia.

And hello again, everyone. I've been popping in and out of the meeting here and there, but it seems that I am the last person before you and the end of the meeting today, so I think that always gets -- I'm lucky to pull that straw every time we meet.

But to start out, one of three measures that we'll be concluding with today -- as you seen on the slides, this is Measure 3638, Care Goal Achievement Following a Total Hip Arthroplasty, or THA, Total Knee Arthroplasty, or TKA. And this can be found on page 12 of the discussion guide.

So just a brief background or description of the measure. This is the percentage of adults patients 18 years and older who have an elective primary THA or TKA during the performance period and who completed both a pre and post-surgical care goal achievement survey and demonstrated a 75 percent or more of the patient's expectations from surgery were or met or exceed.

The patient reported outcome performance measure, or PRO-PM. The score is derived from calculating the differences between presurgical and postsurgical surveys. So a higher score indicates greater care goal achievement. So this is an

outcome measure, but it's a PRO-PM. It uses registry data, claims, electronic health records, instrument-based data, and paper medical records. It is at the level of analysis of the clinician group practice and it's risk-adjusted. It has a statistical risk model with three factors.

I'll move to reliability testing. And you can see on the slide that both reliability and validity pass for this measure. So for reliability testing the developer conducted reliability testing of the patient or encounter level. And the developer tested inter-rater reliability through chart review. Data were obtained from the electronic data warehouse, or EDW, through manual chart review.

The alignment between the manual reviewers and the EDW overall was 97.1 percent agreement with a kappa value of 0.93. The alignment between manual reviewers and the EDW data elements for THA and TKA had 100 percent, so the kappa value there being 1.0, and 94.1 percent agreement, kappa value of 0.87, respectively. So it's THA first and TKA second. The overall agreement between the reviewers and the electronic data warehouse ranged from 89.9 to 99.2 percent.

The developer also conducted reliability testing at the accountable entity level and performed a signal-to-noise ratio approach, and the single to noise ratio generated by the developer was 0.00118 for THA and 0.00004 for TKA. And so some of the SMP members raised concern with the clinician group reliability, noting that the ICCs submits in small sample sizes, and one member acknowledged that the developer mentions that the effect of low sample size in the ICCs meant, however, the SMP member raised concern about -- with the lack of -- between practice variation and the reliability of this measure at the practice level.

Some SMP members mentioned that the reliability testing is sufficient at that patient or encounter

level, yet inadequate at the clinician group practice level due to the sample size, low variability of scores across practice, and no assessment of non-response bias. So that was for reliability.

I'll just summarize validity as well. So again it's not pass. There was no pass on validity and the developer conducted validity testing at the accountable entity level. They did face validity and was assessed with a six-person technical expert panel which convened to provide input on the conditions, groupings, and modeling. And then public commenting was also requested.

The developer indicated that the majority of TEP members agreed that the measure had suitable face validity. And the developer also conducted empirical validity testing which was assessed through measure known-groups and measure-determinant testing.

For the measure known-groups validity testing it was done through a one-question postsurgical satisfaction survey. The developer did not calculate the Pearson correlation due to small sample size. For the measure-discriminant validity, this was tested by comparing the means of care goal achievement, PRO-PM, results by joint for clinician groups with a minimum case volume requirement of at least 25 patients. So the THA adjusted mean was 58.4 percent, and the TKA adjusted mean was 41.3 percent.

Some of the SMP members raised concerns, various concerns really with the empirical validity testing and interpretation due to the small sample sizes overall and the risk adjustment model, testing methodology, apparent homogenous populations, lack of population, variability including social risk, and inclusive -- inconclusive results during measure known- groups testing.

And then SMP member raised concerns whether the risk adjustment model adequately balances priority

decisions for variable inclusions and metrics of fit after model testing. So no evidence for the validation of the risk adjustment model is present.

So that summarizes some of the developer's concerns. You can see the questions up for SMP discussion listed on the slide there. And I'll turn it over to Jack Needleman, who will go through those concerns and areas of discussion.

So, Jack?

Member Needleman: Thanks, Matt.

As is often the case with the NQF staff, they've done a good job with summarizing things. I want to highlight several things from -- that were touched on my Matt just to narrow our focus a little bit.

One is he talked about the use of kappas to measure the reliability of the data elements. That was just done for the exclusions. Does the chart review in the EHR agree? But for the psychometric properties of the instrument itself they had a sample of patients that they did test/retest on and looked at the comparisons there. They said polychoric correlation was a better choice that Spearman, and I don't -- I agree with that. And they picked a threshold of 0.6, which is low, but not extraordinarily low for did they pass. In the pretest group it passed on five of the eight. In the post every one of the eight items were consistent at that 0.6 level.

The real issues with this measure come from the fact that it was tested in an extremely small numbers of sites with an extremely small number of groups. They started out with six sites. Two got excluded from much of the testing because they didn't have enough volume in the case of knee, and three got excluded for hip. So we're looking at data from three or four sites for all the measure at the entity level as opposed to the document level. And I think a lot of the problems with this measure

emerge from that.

We've got a risk adjustment model that has inconsistent results between hip and knee, but even inconsistent results within the three levels that are there for some of the models.

We've got a scatter plot for the validity correlation. They don't do the statistics, but if you eyeball the scatter plot, there's not a lot of variation and there's no correlation between their external measure of post-surgical quality of life and what they're measuring here.

The one other thing I would emphasize as a substantive concern that did not seem to be addressed either in the documentation or in the response was when they talk about missing data, they focus on missing responses on the survey, given that somebody has completed a survey. And those are very low.

What they don't talk about is the precipitous drop between those who took the pre-survey and those who took the post-survey, a drop of more than 50 percent in both hip and knee, and the actual number of cases where they had surveys from both pre and post that enabled them to calculate the measure was 212 for knee and 227 for hip. And there's no discussion of non-response bias in the drop off in who they got post responses from relative to pre responses, whether those would be people who were most satisfied or less satisfied with the surgery.

So it just feels to me like there's a need for much -- this is the issue that Sherrie raised at the beginning of our session. This just feels to me like there's a need for more testing in more places to have better understanding of how this operates at the entity level, not just at the patient level.

Ms. Elliott: Thank you. I do not see any hands raised or comments. Any other subgroup members

have comments or questions?

Member Romano: I can -- hello?

Ms. Elliott: Hi. Is that you, Patrick?

Member Romano: Yes.

Ms. Elliott: Okay.

Member Romano: Oh, there I am. Yes, so I'll be quite frank: I don't see a path to pass this measure on reliability. And I've been studying the algorithm and I think -- of course I don't want to say that the measure is unreliable. I'm simply saying that there isn't the evidence here that would support a passing vote on reliability. Because as we've discussed, the measure score reliability based on the signal-to -noise analysis is -- generates extremely small ICCs, but it's a function of an extremely small sample size.

And the measure -- and theoretically you could go back to data element reliability, but there's no analysis of the data element reliability of the numerator of the PRO-PM. So all of the evidence that's presented is related to the reliability of the denominator and the denominator exclusions but without any information about the reliability of the numerator of this PRO-PM, which is specifically about the -- whether the patient's expectations were met. So we don't have either data element reliability for key data elements or measure score reliability.

So I don't see any path. So I invite my colleagues to explain how you could get to a moderate or higher rating on reliability. I don't see a path.

Ms. Elliott: Any other subgroup members wish to comment or ask any questions?

(No audible response.)

Ms. Elliott: Okay. At this point we can open it up to

the measure developer.

Member Needleman: Wait, wait. I'm sorry.

Ms. Elliott: Oh, go ahead.

Member Needleman: I just want to follow up on Patrick's comment about the small sample size because on page 52 of the discussion guide and the response, or rather on page 52 of the initial submission the measure development team said, and I'm quoting here with one ellipsis, we do not believe the results show poor reliability, but only show that the sample size is small and no conclusions can be reached. That's the developer's statement.

And I think one of the issues for us is at the entity level, which that statement applies to, whether we are prepared to endorse a measure around which no conclusions can be reached.

Ms. Elliott: Okay. Any other subgroup members before we move onto the measure developer?

(No audible response.)

Ms. Elliott: Okay. Does the measure developer have an initial response?

DR. ROZENBLUM: Yes. Thank you and good afternoon. My name is Ronen Rozenblum, and together with Professor David Bates and our team at the Brigham and Women's Hospital and Harvard Medical School we led the development of these care goal achievement PROMs and PRO-PM. We would like to thank the SMP for considering our measure and all the reviewers for their constructive comments. We really found the very helpful.

We would like to briefly point out a few key issues and concerns related to the preliminary analysis done by the SMP and things that just -- people raised just now.

So the first thing is like while we acknowledge an appreciate all the feedback we received from the SMP, we noticed that some of the feedback was specific to the PROMs only, and not the PRO-PM. For example, while some reviewers stated that we had tested our PRO-PM on a group and practice level as specified, others reviewers noted that the measure was only tested on the patient level.

Therefore, we have some concern that some of the feedback about the appropriateness of our PRO-PM testing methodology and outcomes and overall rating about reliability and validity. So basically just acknowledging. And I think it's obvious that the measure discussed here is the PRO-PM and we saw some things that we thought that maybe people were focusing on the PROMs.

The second issue and concern that some people talked about now is about the small sample size, and I think this is the key issue here for the SMP to consider.

Some of the reviewers, as you all heard, raised concern about the small sample size and low variability between the clinician group that were tested. The care goal achievement PRO-PM is a new measure based on a new PROMs developed for this purpose. It's only a new -- it's also a new concept, if you think about it, although a lot of people talk about care goal achievement.

We believe that the outcomes of the reliability and validity testing of the proposed PRO-PM should be assessed in the context of the newly-developed PROMs, not based on already established PROMs widely used in a clinical setting or registry.

Our PRO-PM development constitutes prospective data collection and analysis in real time, real world flow setting, which in our case required collecting PaRIS survey patient data before and after surgery; total hip arthroplasty and total knee arthroplasty, whereas other measures based on existing PROMs

may use a large volume of legacy data and could be tested with a large sample size.

So basically we in order -- because we develop a new PROMs and then the PRO-PM that were based on the PROMs, we had to test it in real environment, which we incorporate our measure into Epic in six sites, six hospitals, and we had to do that prospectively.

Therefore, while we acknowledge the issue related to small sample size and low variability, we feel that testing this new PRO-PM with the largest sample size in another clinician group will delay the use of these valuable PRO-PM by many years. And we're going to touch about it later on when you're going to talk -- ask us specific question.

Based on our comprehensive qualitative interviews with patients, providers and payers; and we had like a few rounds of focus group and interviews with all of them through the years, an environment scale, we believe that there is a real need and great value for the -- a valuable position for this PRO-PM that promote care goal achievement, patient-centered care, and quality of care in orthopedics.

It is also important to mention that there are not currently any PRO-PMs related to care goal achievement in this field. So basically what we ask the committee here; and we will be happy to talk about it further, to take in consideration new PROMs with a new -- no, PRO-PMs that based on new PROM that cannot leverage from existing data, registry or any other data which prospectively will take a lot of time. So what are we doing with that?

Finally, I would like just to acknowledge that our testing show mixed outcomes as mentioned related to reliability and validity. While some of the outcomes were weak, some of them were very --- outcomes we were very pleased to see the reviewers also recognized that.

Just to mention that together with me on this call is Stephanie Singleton, our senior project coding, has a lot of experience with implementation of PROMs, and Dr. Aileen Davis.

Dr. Davis is a professor at the University of Toronto. She has extensive experience in outcomes measure development and evaluation, including PROMs in total hip and total knee. So she led measure development of PROMs and PRO-PM. She also has experience with measure development and submission to NQF and CMS.

And we will all be happy to address the SMP members' questions and concerns. So thank you for everything and consideration.

Ms. Elliott: I'd like to open it up to the full SMP Committee for discussion.

(No audible response.)

Ms. Elliott: David or Christie, I defer to you for next steps.

Member Needleman: Okay. Well, I just want to ask the developer one question, which I said was not addressed in the documentation, at least that I didn't read.

You have this substantial reduction in the number of post-surveys you relative to the number of pre-surveys you have. Was any analysis done of whether the respondents on the post side differed from the mix on the pre side? Any evidence that some people were of the characteristics of the non-responders or the characteristics of the responders relative to the original sample?

DR. DAVIS: So I can start initially on that. One of the challenges with that is no matter how you do that that's going to be biased because a huge piece of the drop in the post sample was simply that people were not sufficiently postsurgery at a

minimum of six months for inclusion. So even if we had looked at that, we wouldn't have known who was a true non-responder versus who hadn't hit the time mark for doing that. So that creates challenges in that. And because it was a new PRO -- PROM and in the context of data collection and with the challenges of COVID we used as much preoperative data as we could.

So, Ronen, I don't know if you have further, but --

DR. ROZENBLUM: Yes, I just want to add something that -- again real-life issues. So, Dr. Needleman, you mentioned that when you're looking at the responsiveness there's two dimension that we should look at that. One of them is missing-ness of items. And because we incorporated into actually a real environment, which from my perspective and --

Member Needleman: Yes, you did a great job there.

DR. ROZENBLUM: -- yes, thank you so much -- which really test the measure, not giving a paper, really tested in a real environment and acceptability and feasibility that I know this is not the method of this committee. But there we had like 0.2 missing data.

Now regarding the data, the non-responders of that, and to your question, one of the issues that we had, because we really did a comprehensive testing; that's what we feel, we had a lot of -- we were surrounded by a lot of experts, that we did our testing mainly in COVID. And how this is related? It's related because when we incorporate into LP, people were able to get PROMs. In our PROMs -- by the way, the PRO-PM and PROMs were imbedded to the HOOS and KOOS with real-life scenario so they could take the survey in patient gateway via patient portal and in clinic.

So what happened, they move resources and we couldn't basically -- people cannot fill out survey in the clinic because of COVID. And later on not all of

them had access to patient gateway or patient portal. And on top of that the cancel lot of surgery.

Why I'm getting into this -- all this information? Because we really wanted to test a true response rate, to your question, that all these issues and noises wouldn't -- didn't give us the opportunity to do that on top of what Aileen just mentioned.

We feel that still based on our observation that the -- we're not expecting, but this is only assumption of mine, a serious or significant response bias here in this measure.

Member Needleman: Thank you.

Matt, am I right in -- for a PRO-PM PROM measure it has to pass the reliability and validity of both the individual item patient level and at the entity level? I thought I saw that on a slide earlier, and that's -- because that's going to affect our potential revoting.

Co-Chair Nerenz: I think that's true.

Dr. Pickering: Yes, that's correct. So that applies for both the reliability testing and the validity testing. Both the data element or the patient level and entity level testing should be done.

DR. ROZENBLUM: Dr. Needleman, can we -- can I ask maybe Aileen to talk little bit about the small sample size and the issues, because we really think this is a critical year when you're developing new measures of PROMs and PRO-PM and you have to do that prospectively for many years to get it about -- I think this is something that the NQF and the committee should discuss, not just about our measure specifically.

Member Needleman: I do and it got raised earlier.

And, Tricia, I actually think you are chairing this, so I'm going to defer to you.

Ms. Elliott: Actually I defer to David and Christie for

that piece.

Co-Chair Nerenz: Yes, well, I'll make another straw person proposal people can take or leave. I've been listening carefully through this for any sort of either developer response or back and forth between members of the subgroup that would suggest a change in views here. We've had a lot of discussion. We've had back and forth. I think actually there's been a lot of useful comment from the SMP to the developers.

My ears haven't heard anything yet that says there's going to be a fundamental change particularly in the favorable direction with a revote. And I acknowledge I'm picking up Patrick's comments specifically about no path.

So I will just propose for others to shoot down that we declare the discussion closed here, but not do a revote because I haven't heard anything that suggests that results are going to change.

Co-Chair Teigland: I agree with David. I did not review this measure, but I've not heard anything I think compelling enough to warrant a revote at this point. So I would recommend that we don't take the time to do that.

Co-Chair Nerenz: And Tricia and Hannah can confirm, if we pull something for discussion; I'm taking that very literally, it means we've committed ourselves to discuss it. It doesn't mean we've then now by the rules of the game committed to a revote. So that's part of why I said what I just said.

Ms. Elliott: Just conferring. Hannah or Elisa, anything to respond to their --

Ms. Munthali: Yes, that's correct. And you posed the question the right way. You just wanted to make sure that everyone was okay with not taking a revote. That's correct, David.

Ms. Elliott: So in terms of next steps with your proposal we'll put any objections?

Member Needleman: I'm not objecting. I do think the developer expressed concern that when you haven't been able to generate enough data to deal with the entity level validation and reliability testing, but you think you've got a solid measure, is that endorsable? That's the same issue that Sherrie raised all the way at the beginning.

I think the answer is no, at least for a measure that has to be reliable at both the patient and the entity level, but -- so I don't think it merits a revote here, but I do think the question needs to be revisited in some other forum to think about this issue of measures that are developed by developers who don't have access to large data sets from the beginning. And but that's I think a conversation for a different time and place.

Co-Chair Nerenz: No, Jack, I agree with your sense, and I think veterans on this panel will remember many discussions we've had about some form of provisional endorsement or something to recognize the circumstance. And we may touch on this a bit tomorrow, sort of a bigger NQF policy issue than we can address.

But we've seen this many times, that a measure that is promising, a measure that has good properties short of early days comes to this either go or no-go statement in NQF and it's not really ready for that final decision, but there's no mechanism by which NQF or our panel can say you're on the right track; keep going, or this seems to have promise. Why don't you see if you can get a few more people to use it? It's maybe tomorrow we will have this or some place in the future.

Member Romano: On additional point. I mean -- and this is sort of directed to the developers. I mean I feel your pain because many of us on this panel are developers in part of our lives. And these

are tough issues in terms of how you get enough of a sample, but our task here is to follow the criteria as they've been set forth. And specifically I would encourage you to think about the data element reliability issue because it's an issue that's not addressed at all in your submission.

You address the denominator but not the numerator, and the numerator here is very important. The numerator is the total number of patients who completed both a pre and post goal achievement survey, who demonstrated that 75 percent or more of the patient's expectations from surgery were met or exceeded.

So we're not given any information about the reliability of that numerator and whether those expectations for example are stable before surgery, are they stable after surgery, those sorts of things.

So I would encourage you to think about presenting additional information, assuming that you may come back to this committee, that really addresses the data element reliability for the numerator as well as the denominator.

DR. DAVIS: So in that --

DR. ROZENBLUM: If I can comment? Yes. Sorry.

DR. DAVIS: Go ahead, Ronen.

DR. ROZENBLUM: No, I was want -- yes. No, Aileen, please go ahead.

DR. DAVIS: So in actual fact we did look at the test/retest and the -- in the pre and the post with the polychorics. And what I also want to point out here, this is a -- in terms of the context of -- for a PRO-PM, this is a formative measure. So it's the individual item reliability that matters. And we moved to the polychoric because we violated assumptions for kappa. And those polychorics by and large were 0.6 and above, which in a polychoric

is considered strong. So the mental health question was low at 0.38, but the other two that were just below 0.6.

So I guess the question I'm asking, just to make sure I understand going forward, is you're actually looking -- that you would like to see the time one, time one pre/post and the time two, time two and done that as a test/retest as opposed to just test/retest within the times? You want it over time? Am I interpreting you correctly?

Member Romano: Well, I think so. I invite Sherrie's input as well, but the measure that we're asked to evaluate is whether 75 percent of the presurgical care goals were satisfied or exceeded. So that's a very tricky concept that involves information from both the pre and the post-surveys. And that itself has measurement properties that need to be considered and explored separate from the measurement properties of the pre-survey in isolation and the post-survey in isolation.

DR. DAVIS: Okay.

DR. ROZENBLUM: If I may say something? I appreciate your comment and I don't know if I'm addressing your concern, but let me tell you what we did, okay, in terms of data element.

We actually -- all the measures, all the alignment of the measures that we did in this testing was base on measures on pair data that completely aligned with our measure-specific agents. So they met the numerator and denominator. Okay?

So I'm saying it carefully because I don't know if you would like us to do something specifically for the numerator, but each one of the cases that we assessed, the two people individually assessed, were in line with the complete PRO-PM, met the numerator and the denominator. I believe that we did what you're saying, but maybe I don't understand completely that you want us to do that

separately. But just for you to know, all of them met also the numerator. So we actually chose only the cases that actually were in line with the measure specification, including the numerator. I'm not sure if I addressed your concern, but that's what we did actually.

Ms. Elliott: Sherrie, you had your hand raised?

Member Kaplan: Well, just to respond to Patrick's issue, I think test -- for somebody like me test/retest reliability means over time. It means for the same sample do you get the same result, where you wouldn't expect true score variation? You would expect just a replication of a person's or an entity's original score. So test/retest for me implies time. And what half reliability means, if you take a sample and split it in half at a single point in time, you get the same answer for random halves of that sample. So there's little terminology maybe confusion, but that's what people like me mean.

Member Needleman: I want to give the developer here credit where credit is due, and there's a lot of credit due. They found the sample, they asked people, a sub-sample of the folks who did the pretest, to redo the test. And that's what they've compared.And then for the folks after surgery, they found the sample that did the post-test and they asked them to do the post-test again, and that's what they compared. And so that's been done.

We heard Eleanor --

DR. DAVIS: Aileen. Yes.

Member Needleman: -- Aileen, thank you, describe the case -- the three cases where they didn't hit the 0.6 coefficient, and two of them were pretty damn close. So Patrick then said now we've got to look at how stable the score is, which involves looking at the presurgery/postsurgery calculation.

And I'm not quite sure how to do that sitting here,

but it might be calculating it the four ways you can calculate it using the two different pretest scores and the two different post-test scores, postsurgery scores and seeing how stable the results are there. And that's there. And I thought that's what I heard Patrick asking for some analysis to show how stable the scores are.

DR. DAVIS: Yes, and that's what I was clarifying that he intended. Yes.

DR. ROZENBLUM: No, the only thing that I would like to add just from my perspective -- I'm looking at the PRO-PM, I'm looking on the screen because I have the results -- we did two testing on the reliability side, the data element and then the signal-to-noise. The data element I'm sure -- there is concern, but we think that it show very high alignment. And obviously because of the small sample size we had a poor basically signal-to-noise.

In terms of validity testing for the PRO-PM we did three testing. Face validity, which was showed 100 percent agree with the TEP. Then measure discriminate validity where we assumed that with -- looking at the scoring of the PRO-PM that hip patients, total hip patients, that that replacement will have a better score because the recovery will get a better score than the knee. And we saw that. So we got also that marked.

And then there's a measure known-group validity. Obviously again we couldn't calculate the person correlation because of the small sample size.

Just want to prove to the committee, and I know that you know that, that actually there is -- and I said at the beginning, there's actually mixed method. And I know maybe the most important outcomes that you're looking didn't show significance because of the small sample size, but going back again there's a paper Aileen can talk about that -- a general paper that's suggesting that we will need 45, you know, all those sites to do that

prospectively. And in a measure like -- that promoting person-centered care.

So we going to talk with you guys -- it's not going to be next year to do that, even if we would like to do that. And this is what I'm doing for my living, person-centered care. It's like it's going to be five years for now where we're going to have 45 sites because we have to implement it prospectively. And I know this is another maybe topic for consideration, another topic, but that's what we feel.

Ms. Elliott: Okay. Great, great. Any other comments or discussion from the SMP?

(No audible response.)

Ms. Elliott: Okay. I do not see any hands raised or chat.

So, David, can I circle back to you to wrap up this measure?

Co-Chair Nerenz: I think we have. I mean there's plenty of discussion. And I think all of us feel great sympathy for the challenge the developer faces here to get this out in large enough use to have a large data set with which to do many of these analyses. It's kind of a chicken and egg problem. I don't know that NQF endorsement would actually have a great deal with that expansion, but it's a challenge. But this may be something off line where either individual SMP members or the group could offer even more specific suggestions than we have. But I think we've kind of reached the end point at least in terms of the decision we have in front of us.

So with that my suggestion is we move on to whatever we best need to move on. I understand we've now got a short time block and still need to work in a couple things with our public comment. So I'll defer to staff on how best to juggle our remaining agenda.

Ms. Elliott: Okay. So to confirm, on this measure we're not opening it for revote as you recommended earlier. Correct, David?

Co-Chair Nerenz: Well, I'm the only one who is talking. Others who -- if there's another strong wish to revote, chat, hand up. Now is the time.

Ms. Elliott: I am not seeing any hands raised or objections. And you had asked for clarification earlier, so I think we're good.

Given the time, if we started a discussion on a measure, I don't think that we would have enough time to complete the discussion, so I would like to propose the two remaining measure, which are 3639 and 3667, be moved to tomorrow, Wednesday's meeting. And we would also revisit the one measure that was proposed for a revote and moved to -- or reconsideration and moved to tomorrow as well. So there would be three measures on the agenda for tomorrow.

Any discussion warranted on that proposal or input?

3639 Clinician-Level and Clinician Group-Level Total Hip Arthroplasty and/or Total Knee Arthroplasty (THA and TKA) Patient-Reported Outcome-Based Performance Measure (PRO-PM)

Member Romano: Could I ask, since we have developers here, I assume, 3639 I think was a pass/pass. It might be a brief discussion if we --

Ms. Elliott: Okay.

Member Romano: -- discuss it. I don't know -- I don't recall what the issue was that prompted it to be pulled, but maybe we could at least clear that one.

Ms. Elliott: Okay. Totally open to that suggestion.

So maybe we can bring up that next slide, which is there.

Matt, do you want -- I think this is your measure as well. Do you want to give an overview and maybe we could squeeze this one in today?

We do need to leave at least 10 to 15 minutes for public comment. So we'll maybe kick this off and see how far we can get based on Patrick's suggestion.

Dr. Pickering: Sure. Yes, I'll go ahead and kick this off.

So this was 3639. It's the Clinician-Level and Clinician Group-Level Total Hip Arthroplasty, or THA, and Total Knee Arthroplasty, TKA, PRO-PM.

So I won't really go too much into detail around the description of the measure, but it is an outcome PRO-PM measure using claims and instrument-based data. It's at the clinician group and individual clinician level. It is risk-adjusted with 19 factors included in it. You can see it's pass on both reliability

and validity.

One of the areas of discussion is related to reliability testing in which the patient-level or encounter-level testing for reliability -- the developer did not conduct patient or encounter-level testing for PRO-PM in the specified measure population, time frame, and setting required. Rather, they used a test/retest and internal consistency to assess reliability for both PRO-PM instruments or PROMs, patient reported outcomes; i.e., HOOS Jr. and KOOS Jr.

So the internal consistency was calculated using a Pearson separation index. For both instruments internal consistency ranged for 0.4 to 0.87, and interclass instruments -- or excuse me, interclass correlations for liability were between four dimensions: pain symptoms, activities of daily living, sport and recreation function, and quality of life, of the HOOS Jr. and KOOS Jr. with range of

0.75 to 0.97.

So reliability testing was also conducted at the accountability entity level. Signal-to-noise approach was used.

Among the clinician, individual clinician and clinician groups with 5 in 10 case the signal-to-noise ratio yielded a median reliability score ranging from 0 to 0.79 and 0.79 to 0.85, respectively. And then the mean was 0.69 for clinicians with at least five cases.

And then for the clinician and clinician groups with 25 cases the signal-to-noise ratio was a bit higher -- higher median reliability scores ranging from 0.79 to 0.97 and then 0.79 to 0.99, respectively.

So one SMP member raised concern regarding the variation responses as it relates to social risk and the experiences among racial groups may be underrepresented in the sample.

They also conducted for validity. They also did patient-level and encounter-level testing. The developer evaluated responsiveness for both instruments using standardized response means and compared against two other previously validated PROMs. The correlations ranged from 0.84 to 0.94 for HOOS Jr. testing correlations where the KOOS Jr. testing ranged from 0.72 to 0.91.

I'll just sort of mention that one SMP member raised concern with a potential for measurable improvement related to the floor and ceiling effect. For the HOOS Jr. testing the ceiling effect did not meet the 22 points to support substantial clinical benefit.

Validity testing was also conducted at the accountable entity level and face validity was assessed by asking a 17-member TEP to respond to statements using the six-point scale. Seventy-six percent either strongly or moderately agreed with the statement that this measure as specified will

provide valid assessment for improvement in functional status and pain following an elective primary THA or TKA.

Some SMP members expressed interest in observing descriptive statistics for those patients with no response to allow for construction of models to adjust non-response prior to assessing reliability. And then several SMP members raised concern with non-response bias and the accuracy of developer's validity assessment as 37 percent of the sample was excluded due to missing PRO scores, or patient reported outcome scores, 10 percent due to missing factors and 2 percent without clinician attribution. But again, it did pass on validity. It passed on reliability, and I mentioned reliability being the area of discussion.

I will just remind the SMP that it is both at the data element level and the accountable entity level for patient reported outcome performance measures that we would want to see testing or a demonstration of reliability and validity at both levels. According to our criteria for reliability testing specifically, prior evidence of reliability of data elements for the data types specified in the measure can be used as evidence for those data elements. Prior evidence could be published or unpublished testing that includes the same data elements. It uses the same data type, so chart abstraction claims, and is conducted on a sample as described previously. So just letting the SMP know that for reliability testing previous testing or data can be used and that can be pulled from those sources that I mentioned previously.

So with that summary I will also note that the developer is Yale CORE, who's on the call; measure steward being CMS. And all of this can be found on page 14 of the discussion guide. And you can see the questions listed here. And then I'll turn over to our lead discussant Daniel to see if there's any discussion from the SMP.

Daniel?

Member Deutscher: Thanks, Matt. Thanks for this wonderful summary again.

First of all I'll start by saying that I did not ask that this measure be pulled for discussion. I think the vote is fine. I did vote on the validity an insufficient vote and I wanted to explain why and maybe through this measure raise a little bit more of a general issue that we may decide to discuss maybe at a different time.

So my main concern was about the accountable entity-level validity. I did not have important enough concerns related to the other comments that you mentioned that would, at least in my view, make me not pass the measure. So I'm going to briefly review that concern and then I'll pass it onto my colleagues to raise the other concerns if they want to do that.

So as you mentioned, this is a PRO-PM and it needs to reliable and valid on both levels. This measure is similar to an endorsed measure; I think it's 3559, which is basically the same measure at the hospital level. Now given the fact that the accountable entity level was really what is new about this measure compared to the other measure I did rate validity as insufficient because validity at the accountable entity level was not tested other than the measure's face validity presumably at that level.

I want to also review briefly the criteria of NQF, and NQF specifies that for instrument-based measures including PRO-PMs validity should be demonstrated as you said for both the instrument and the computed performance score. However, face validity testing of the computed measure score is accepted at initial endorsement and validation. So I'd like to review briefly the face validity that was done here at the entity level.

And it was assessed basically using two questions.

The questions were directed to a technical expert panel I think of 17 people and basically 4 patients. The questions were about the level of agreement with two statements. And as you mentioned there were six response categories and they started from strongly agree. And it was followed by moderately agree and all the way to strongly disagree. I'd like to review these questions in order to raise this maybe a little bit more basic issue.

So the first question was the clinician and clinician group level, total hip or total knee PRO-PM, as specified; so as specified by the level of the measure, will provide a valid assessment of improvement in functional status and pain following elective primary surgery. I interpret this question as being focused mainly on the improvement in functional status.

I'm going to state the results a little bit differently compared to what the developers did and I'm going to focus on the first response category, which is strongly agree. Seven of the seventeen technical expert panel members; that's about 40 percent, responded strongly agree. Two of the four patients also responded strongly agree, so fifty percent.

The second question that was used was similar: The measure as specified, can it be used to distinguish between better and worse quality care among clinicians and clinician groups? I interpret this question as being more focused on the accountable entity level, at least compared to the first question, which I interpreted as being focused on functional status and pain improvement.

The results here, and again focusing on that first category, were that only 3 of the 17 technical expert panel members; so that's less than 20 percent, responded strongly agree. The patient scores were the same. Maybe not the same patients. But two of the four patients responded strongly agree.

The developer's interpretation of these results -- the interpretation was that the -- and I quote, the vast majority of the technical expert panel and patients endorsed the face validity of this measure as demonstrated by the widespread agreement in responses to the two face validity statements. So the main difference between face validity of the data element or the entity level is really the specification of the level being asked about using this question.

Two issues here: First, I to some extent may question the ability of a person to truly, or maybe we should say validly be able to differentiate between these two levels when answering these types of questions. So if I'm asked a question can I really be able to differentiate the ability of a measure to distinguish or to identify change or improvement for a patient compared to an accountable entity level? Maybe yes; maybe no. I'm not sure. And I think that this type of validity at the entity level would be at best maybe able to support some kind of soft evidence of validity.

I also, as I noted, have some concerns about the interpretation of the results from these two questions as presented by the developers. The interpretation of face validity is many times a little bit arbitrary. That's not uncommon. But here I did raise some concern and I'd be interested to hear the developer's response, although again I don't think that's enough given the existing s requirements of NQF. It's not enough of a reason to fail the measure.

So a more general clarification question that I have for ourselves and for the NQF staff is that in this case where we specifically say that a PRO-PM needs to be tested on both levels for both validity and reliability, could it be that the face -- allowing for just family-centered at the accountable entity level, may that be an easy way out of that specific demand to the PRO-PM? Or in other words, was this

the intention when initially differentiating between these more complex measure and other measures when asking for these specific requirements?

So that's my main question. It's maybe a more general question that we could decide to discuss maybe at another time given our limited time today. And these were my comments basically and the reason why I thought the validity testing was insufficient. I'll pass this onto my colleagues for any other comments. Thank you.

Ms. Elliott: Sherrie, you had a comment or a question?

3649e Risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) electronic clinical quality measure (eCQM) (BWH)

Member Kaplan: This is the measure that I started -- I didn't draw a line between my comments on this measure and 3649e, and so I kept reading on down. So I actually started commenting, as was accurately pointed out by Patrick and David, on this measure.

But I had some concerns about this measure because of the ceiling effect problems and their decision to make this at the -- a cut point for a clinically-significant difference of 22 points, which is a whole standard deviation, as noted by Lyman and Lee, which is not what they cited in that article. So I had some concerns about whether or not this was actually going to help -- be useful for distinguishing at the practice level.

The other concern I have is these are really high ICCs at the practice level. And that raised issues for me about how that was computed because what you're looking for in ICCs for this entity is the between versus with -- over between plus within-practice variation, and that includes in the nearer term the within-patient across items in a multi-item

measure like this.

So that's a really high ICC and my concern is that was done in a way that I don't understand. Because I would have expected it to be as for CAP measures and other things down in the range where we've been seeing 0.05, 0.0 -- at the highest 0.1. And it raised issues for me about how that was done, so I had some concerns.

But again, because my colleagues don't -- I mean, I'm going to turn it back to you and David to figure out whether -- since this is passed whether that warrants further discussion.

Ms. Elliott: You're on mute, Patrick.

Member Romano: I think many of us share this concern about the ceiling effect, so we should hear the developers and then find out -- consider whether these concerns are grave enough to support a revote in combination with the face validity concerns that Dr. Deutscher has mentioned.

So again to put a little context on this, the ceiling effects are reported in 37 to 46 percent of patients in HOOS, 19 to 22 percent of patients in KOOS. And because the developers are looking at the difference, the patient-level difference, and they're evaluating that relative to an SCB, a certain magnitude of improvement -- so basically they're looking for a magnitude -- I think was it 22 points, of improvement. The standard deviation, at least in one of the prior papers, I found 1413. So perhaps even one-and-a-half standard deviations.

So a fairly substantial percentage of people could be in the range preoperatively where they can't meet the target even if they feel perfect after the surgery. So this is a critical question to understand what happens to people who are feeling pretty good before surgery, but bad enough that they really want to have surgery. And then they get the surgery and they feel wonderful after the surgery.

They've had a fantastic outcome, but yet they don't meet the SCB threshold. So this is a question I think many of us are interested in.

DR. SUTER: This is Lisa Suter from Yale CORE. Is this an appropriate time for the measure developer to respond? And can you hear me?

Ms. Elliott: Yes, we can hear you and go ahead and respond. Thank you.

DR. SUTER: Great. Thank you.

I've heard a lot of different questions. I think Dr. Romano's last question seems to be what we're focused on responding to, which is the issues raised around ceiling effects.

So I think the one thing that may not be evident to this group but our test and clinical working group and patients felt strongly about, more so the clinicians, is that the way we have created this measure actually not only ensures that many patients who in other circumstances, or if you were focused on the end result of the measure, might be passed over or may have reduced access due to this measure as an unintended consequence because they start off with very poor scores and therefore will not achieve an end score that might -- if you used an end score to -- threshold to define the measure's success.

Similarly, a number of orthopedics feel that there has been scope creep for hip and knee replacements. And therefore on the opposite end of the symptom spectrum there are patients who have mild symptoms who really ought to be managed non-surgically with other modalities before having them undergo hip and knee replacement. Because there is approximately a one percent linear trend per year of hip and knee replacement recipients who need revision surgery, even within a year of having the surgery.

So from an orthopedist's perspective; I'm a rheumatologist, not a orthopedist, the ceiling effect is not concerning to them because they actually like the fact that you really needed to have substantive symptoms before surgery in order to see a benefit from surgery out of this measure, and they consider that as a way to reduce inappropriate use of hip and knee replacement.

So that was specifically addressed and the patients verified the empiric results that were -- that suggested the substantive clinical benefit. They felt that that space was a reasonable delta to focus on and represented both something that was worth having surgery for and would represent improvement in their symptoms. And this was gone through with them question by question showing them what differences in scores might -- how they might experience that based on the questions. So it was done in a very detailed way.

I hope that addresses the concerns about ceiling effect. It's a clinical response and an implementation or policy response, not necessarily a methodologic one, but I think very helpful for you to understand.

Member Romano: So just to be clear, you are explicitly failing any patient who was over 78 on the total hip measure or over 80 on the total knee measure preoperatively? You're explicitly calling that a failure because that patient shouldn't have had a total hip arthroplasty? Is that right?

DR. SUTER: So realize that these are risk-adjusted measures. So to the extent that there are going to be cases around the margins where different risk factors might change that exact cut off, yes, the concept was that people who have very minimal symptoms; this is a 0 to 100 score with 100 being absolutely perfect help, no symptoms -- but yes, the feeling was that these individuals shouldn't be going through such a significant and potentially

complicating surgery.

Member Romano: And could you address Dr. Deutscher's comment about what appears to be marginal face validity on the second question?

(No audible response.)

Member Romano: We usually ask developers to explain when there's substantial disagreement in the TEP, why there was substantial disagreement.

DR. SUTER: So sorry. We had 14 TEP members and 4 patient working group members endorse that it distinguish quality. I'm not really sure I understand the question that it was marginal in the second face validity question.

I will say that related to those questions they were used for the hospital measure that was passed. They've been used by other developers and in other measures. We think that they're helpful to address face validity.

And we've been asked by NQF staff to use specifics questions that present you with a specific result. We combine strongly agree and agree. If NQF is concerned about combining strongly agree and agree as measuring face validity, it would -- we will respond accordingly in the future.

Member Deutscher: Yes, I'd like to comment on that. I don't think it was strongly agree and agree. I think it was strongly agree and moderately agree. And since we're -- in the first place we're doing -- we're really doing here looking at face validity instead of actual validity testing at the accountable entity level. So we're kind of giving, as I said, maybe a little bit of an easy out of that specific requirement for PRO-PMs.

But, so when we do that at least I think it would be reasonable to expect strong support. And if I look at the response categories, I would -- at least I could

come up with an argument saying that what I really want to count is those that really agree with the statement now, and not those that also moderately agree. So does moderately agree support face validity or not? Obviously that could be argued. And I don't know if you want to get into that now because I could come up with a reasoning in both directions.

But I did want to raise the overall -- I don't think this is convincing evidence that we can support the face validity of the accountable entity measure. I also think there is a difference between the two questions. You used one as more oriented towards the functional improvement. The other one is really more oriented into looking at the entire level, and the results were not as good for that second question compared to the first one. That also to my view does not give strong support to face validity at the accountable entity level.

Member Romano: And, Dr. Suter, just to --

DR. SUTER: Thank you.

Member Romano: -- clarify, so NQF, the guidance page 21, says the degree of consensus in any areas of disagreement must be provided, discussed. Two individuals disagree. So we just need to know why did they disagree? What were the factors that were driving your difficulty getting support?

DR. SUTER: So I think in general the concern -- any concerns articulated about this measure and the concern for not supporting the measure was that despite the fact that this measure was created out of robust data from a number of clinician and clinician groups, it was -- I mean it was collected through a prospective incentivized voluntary reporting program which substantiates its feasibility.

There were concerns that because of the way the measure was incentivized and only 50 percent the hospitals in the comprehensive joint replacement

model in CMMI, which was the model in which this data was collected -- those hospitals who submitted the data were only incentivized at the beginning to capture 50 cases or 50 percent and that those thresholds for response were increased over time.

And there have been some concerns among people including the two TEP members that they would like to see the measure in use and in a broader or different data sample. And that I think if you asked them, they would indicate that the best way to do that is to have the measure in use in a voluntary way in order to collect additional data so that they could feel more confident in those results.

All the remaining members of the TEP and the patients felt comfortable with the measure as currently specific despite those limitations of the original data set. And I'm happy to -- I mean I don't -- I think in terms of what level of agree should be considered meeting criteria, I really defer to the NQF staff since in prior measures we have not had the responses discussed quite this way.

Dr. Pickering: This is Matt. I know that we're really getting close to a hard stop here, and I appreciate David Nerenz' comment in the chat box.

Maybe before we go to that I'll just state to Daniel your question about face validity is an acceptable form for new measures for NQF. It is looking at the measure score level or at the accountable entity level.

And to what degree is sufficient as far as agreement? That in our criteria is a question that the evaluator; in this case the SMP, needs to make a decision on whether or not you agree that there's substantial agreement being placed on that question of whether or not this measure can discern good versus poor quality with the information that the developer has submitted to you, which is what we've been discussing.

There is that element of patient level or the encounter level for PRO-PMs. This can be established again by previous evidence that's been submitted in which the developer did provide a response to those concerns related to some of the previous validity testing results that they've submitted with this measure. So that's taken into consideration as well.

But I just wanted to touch on those two points for consideration as David Nerenz has mentioned maybe holding this thought until we pick it up again tomorrow. So with that maybe I'll turn it over to David or Tricia as we have four minutes left.

Ms. Elliott: Thank you, Matt.

And, Patrick, you added a comment in the chat. If you could share that real quick?

Member Romano: Yes, I'm just pointing out that I'm not moving to revote. I think this was -- the vote is here. I don't know if any people changed their mind. We've had a robust discussion. So I leave it to the chairs.

Ms. Elliott: David or Christie, any comments?

Co-Chair Nerenz: I'm just watching the chat. It would be nice if we could close this out without a revote, taking that as a motion. I'm watching the chat. Let's see what people think. And it's the subgroup's decision.

Ms. Elliott: We have so far three comments, four comments, five comments for no vote. They can in rapid fire there, so let's see. One, two, three --

Co-Chair Nerenz: That's what we need. Going once, going twice --

Co-Chair Teigland: So if someone really wanted to revote they would raise their hand quickly, right? They know time is short.

Co-Chair Nerenz: I think the clock just ran out. Let's close this one out.

Co-Chair Teigland: Yes.

Co-Chair Nerenz: Good discussion. Move on.

Ms. Elliott: Excellent. Okay. Thank you so much.

At this point I would like to open up the meeting for public comment. And just a reminder to either raise your hand, feel free to chat. Or if you're just a call-in phone user, feel free to speak up as we open it up for public comment.

(No audible response.)

Ms. Elliott: Okay. I'm checking with the team. Any hands raised? I'm not seeing any new chats. Just pause for another moment.

(Pause.)

Ms. Elliott: Okay. So I do not see any hands raised, I did not hear anyone come off mute, and there's no new chat messages.

Hannah, we were going to hand things over to you to -- on the next steps, but I think our next steps have changed a little bit based on the discussions today. Do you want to go ahead and maybe state where we're headed for tomorrow?

Ms. Ingber: Yes, I can do that.

If you'd go to the next slide, Gabby?

Just a reminder that tomorrow's meeting will be from 3:00 to 5:00 p.m. Eastern Time, but we'll first start with our discussion of 3667 and then move to 3652 after the panel members have had a chance to review the materials as we discussed before. And then as time allows we will give a CSAC update and discuss the reliability thresholds tables and the maintenance reliability and validity testing at the

accountable entity requirements that we've been discussing in past SMP non-measure evaluation meetings.

Next slide, please? I'll hand it back to you, Tricia.

Ms. Elliott: Okay. Great. Thanks so much, Hannah.

David or Christie, any closing comments before we wrap up for today?

Co-Chair Nerenz: As always thanks to everyone for diligent thought, dedication, good suggestions, respectful comment back and forth. Always good.

Co-Chair Teigland: Yes, great discussions today and I think we have our next SMP meeting agenda pretty well set here. We've got some real important things to discuss, so we'll look forward to that as well. Thanks, everyone.

Ms. Elliott: Thank you both.

And thank you to all the committee members for their time and dedication today and to our CMS colleagues who joined the call as well as the measure developers. We appreciate the dialog and the additional inputs and we'll see everyone tomorrow at 3:00 p.m., Eastern Time. Thanks again.

Co-Chair Teigland: Thank you, SMP Team. Great job.

Ms. Elliott: Thanks, Christie.

Co-Chair Nerenz: Thank you. Bye-bye.

Co-Chair Teigland: Bye.

(Whereupon, the above-entitled matter went off the record at 5:01 p.m.)