National Quality Forum

+ + + + +

Cost and Efficiency Standing Committee

+ + + + +

Spring 2022 Measure Evaluation Meeting

+ + + + +

Tuesday

July 12, 2022

+ + + + +

The Committee met via Videoconference, at 10:00 a.m. EDT, Sunny Jhamnani and Kristine Martin Anderson, Co-Chairs, presiding.

Members Present:

Sunny Jhamnani, MD, Central Arizona Heart Specialists, Co-Chair
Kristine Martin Anderson, MBA, Booz Allen Hamilton, Co-Chair
Bijan Borah, MSc, PhD, Mayo Clinic
Cory Byrd, Humana, Incorporated
Amy Chin, MS, Hospital for Special Surgery
Risha Gidwani, DrPH, RAND Corporation and UCLA School of Public Health
Emma Hoo, Purchaser Business Group on Health
Sean Hopkins, BS, New Jersey Hospital Association
Pamela Roberts, PhD, OTR/L, SCFES, FAOTA, CPHQ, FNAP, FACRM, Cedars-Sinai
Danny van Leeuwen, Opa, RN, MPH, Health Hats

NQF Staff Present:

Taroon Amin, PhD, Consultant to NQF
Poonam Bal, MHSA, Sr. Director
Matilda Epstein, MPH, Associate
Victoria Quinones, AA, PMP, Project Manager
Isaac Sakyi, MSGH, Manager
Tristan Wind, BS, ACHE-SA, Analyst
Leeann White, MS, BSN, Director

Also Present:

Sam Bounds, Acumen
Rose do, Acumen
Joyce Lam, Acumen
Heather Litvinoff, Acumen
Sri Nagavarapu, Acumen

Contents

Proceedings

(10:01 a.m.)

Welcome and Review of Meeting Objectives

Ms. White: I want to welcome everyone to the Spring 2022 Cost and Efficiency Measure Evaluation meeting. Greetings and good morning. My name is LeeAnn White, and I'm the director supporting the Cost and Efficiency project team for the spring 2022 cycle.

First, I do want to thank all of you for your time and participation this morning. I do understand that it is a significant amount of time and effort to review the measures and prepare for today's meeting.

But I'd also like to extend a thank you to the developers for being on the call today as well. We do also recognize the significant time and effort that goes into the testing, preparation of the materials and measures submission. I do want to highlight those efforts and thank them for their team as well.

Lastly, I do appreciate your continued patience and understanding as we continue to meet virtually in the pandemic. We do understand the challenges that do accompany virtual calls, and we do look forward when we can convene in person.

However, in the meantime, we do appreciate your understanding and thank you for your continued support. We have learned some tips and tricks along the way to help bridge some of those virtual gaps. Hopefully, those work on our web meeting today.

If you give just a moment, I will have our team member pull up our slide deck. I do see some attendees still dialing in. We'll just pause here for a moment to get our slide deck pulled up. Welcome to everyone who's joining the call. Good morning.

Okay, I will hand it over now to our co-chairs, Kristine Martin Anderson and Sunny Jhamnani, if you'd give

the welcoming remarks to the Standing Committee this morning.

Chair Martin Anderson: Hi, everyone. It's Kristine. Thank you again for joining. I'm looking forward to what should be a good discussion. Thank you.

Chair Jhamnani: Good morning, everyone. Sunny here. Thank you for being present here. Let's have a good discussion on the measures. Let us know if you have any questions or we can help you in any way.

Ms. White: Wonderful. Thank you, Sunny. Thank you, Kristine.

Next slide, please. And next slide, please. All right, I'd like to take a brief moment to quickly review a couple of housekeeping reminders. As most of you know, we are using the Webex platform to host this measure evaluation meeting today.

If you are having any technical difficulties, please reach out to our team. We're here to help you, assist you with your audio and your visual, any troubleshooting that you may need for us to address.

Our team is also standing by via chat, and you can email us directly at efficiency@qualityforum.org. We will be monitoring our project inbox throughout the call today.

In the spirit of engagement and collaboration, I encourage us all to use our video so that we can see each other's faces and bridge some of those virtual gaps.

If you're not actively speaking, we do ask that you please place yourself on mute to minimize any background noise and interruptions. To mute and unmute, there is a microphone icon located at the bottom of your screen. We highly encourage everyone to use the chat box feature and raise hand feature throughout the meeting today as well.

Raising your hand does alert the host, and a raised

hand icon does appear in your video and at the participants panel. Our team will be monitoring for raised hands throughout the call.

To raise your hand, you can do this in a couple of ways. At the bottom of your screen, you'll notice a hand icon. If you click on that, it will raise your hand and then clicking on that icon again will lower your hand.

You can also scroll to the participants list and find your name, and there's also a raised hand icon that will appear when you do find your name.

Once the meeting begins, our senior director of Measurement Science and Application, Dr. Matt Pickering, will conduct roll calls and review disclosures of interest. It is important to note that we are a voting body, and therefore we do need to establish a quorum to vote on our meeting today.

If you do need to step away from the call, please let us know, the NQF staff team know. We do need to monitor attendance throughout the meeting today. We do ask you to let us know when you leave and then upon your return.

Next slide, please. I will go ahead and introduce our team. Again, my name is LeeAnn White. I am the director supporting this team for the Cost and Efficiency Project. Our manager is Isaac Sakyi. Analyst is Tristan Wind. Our associate is Matilda Epstein. Our senior directors are Poonam Ball and Dr. Matt Pickering. Our project manager is Victoria Quinones. And our consultant is Dr. Taroon Amin.

Next slide, please. I will touch on some of the agenda items that we have listed here and what we'll be covering today. We will begin by conducting roll call and disclosures of interest.

Two disclosure forms were sent to you. One is our annual disclosure form, and then the second is specifically related to the measures we are reviewing. So we refer to that disclosure as the measure specific

disclosure of interest form. We must receive both of those forms from you to review any potential conflicts.

If we do not receive those forms, unfortunately, you will not be able to participate in the discussions or the voting today. Our team members will reach out to the Standing Committee members that we still have disclosures outstanding. We do ask that you please fill out those forms promptly and return them back to us so that you are able to participate on a call.

After we complete disclosures of interest, Isaac will provide an overview of the evaluation and voting process. Tristan will conduct a voting test.

We did send a Poll Everywhere link to everyone's email at approximately 9:45 Eastern Time this morning. If you can please check your inbox, it will contain a Poll Everywhere link for live voting. If we meet voting quorum, we'll be using that on the call today.

After the voting test, there will be a brief introduction of the measures under review, and then we'll hand the discussions over to our co-chairs to facilitate the discussions. Within the discussion is each criterion, and we vote on each criterion.

We also did want to notify that NQF has created a designated timeframe specifically for developers to respond to any questions or concerns that is brought up by the Standing Committee throughout the discussions, and they'll provide that time to provide their clarifications.

The co-chair and staff will collect any questions from the developer during the discussion for each criterion, and then we'll open that time up prior to the vote so that the developer can be given the opportunity to respond to those questions and clarify any information prior to the Standing Committee vote.

The last vote will be an overall recommendation for

endorsement for the measure. Today, we will not have a related and competing discussion as there were no measures identified as related or competing for any of the measures under review today.

Following the measure discussions, we will host an opportunity for NQF member and public commenting, and then we'll conclude with next steps and what to expect moving forward.

Next slide, please. Now, I will hand it over to Dr. Matt Pickering, who will review disclosures of interest and conducted roll call.

Matt?

Dr. Pickering: Can you hear me okay?

Ms. White: We can, yes.

Introductions and Disclosures of Interest

Dr. Pickering: Thank you. All right. Well, hello, everyone. It's good to see some of you again. Thank you as always for your time and your participation in this work.

As LeeAnn had mentioned today, we'll be combining introductions with disclosures of interest. You did receive those two disclosure of interest forms. One is our annual like LeeAnn had mentioned, which we do every year. And another one is specific to the measures that will be discussed under this review cycle.

So 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 and the work that we'll be doing today. We are especially interested in grants, research or consulting related to this committee's work.

Just a few reminders, you sit on this group as an individual. You do not represent the interest of your

employer or anyone who may have nominated you for this committee. We are interested in any disclosures of both paid and unpaid activities that are relevant to the work in front of you.

Finally, just because you disclosed does not mean that you have a conflict of interest. We do verbal disclosures in the spirit of openness and transparency. As I go down the list of names, if we do not have one of the disclosure of interest forms, whether it be the annual or measure-specific, I'll just recognize that and then the team will send an email to you just to make sure that we get that from you before we go into the measure discussions just because you won't be able to participate in those nor be able to vote on the call if we have quorum.

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

If you do not have any disclosures, please just state, I have nothing to disclose to keep the conversation moving. If you experience trouble unmuting yourself, please raise your hand so that your staff can assist you.

So I'll go down the name. And I apologize as well if I mispronounce any of your names as they appear on the slide. I'll start at the top. Sunny Jhamnani?

Chair Jhamnani: Good morning. Present. No disclosures.

Dr. Pickering: Sunny, would you mind just stating the organization you're representing -- or not representing, but with?

Chair Jhamnani: Private cardiology practice.

Dr. Pickering: Great. Thank you, Sonny.

Kristine Martin Anderson?

Chair Martin Anderson: Kristin Martin Anderson, Booz

Allen Hamilton. Nothing to disclose.

Dr. Pickering: Thank you.

Robert Bailey? Robert Bailey?

Okay. Bijan Borah? Bijan Borah?

Member Borah: Sorry, hi. I was muted. Yes, hi. I'm Bijan Borah from Mayo Clinic. I am a consultant to Exact Scientists and Boehringer Ingelheim, but nothing to related to the work with this committee work. Thank you.

Dr. Pickering: Thank you so much, Bijan. Cory Byrd?

Member Byrd: Good morning. This is Cory Byrd. I'm with Humana, Incorporated. Nothing to disclose.

Dr. Pickering: Thank you, Cory. Amy Chin?

Member Chin: Hi. Amy Chin. I'm with the Hospital for Special Surgery. Nothing to disclose.

Dr. Pickering: Thank you, Amy.

Lindsay Erickson? Lindsay Erickson?

Okay. Risha Gidwani?

Member Gidwani: Hi, good morning. I am Risha Gidwani with the RAND Corporation and the UCLA School of Public Health. I have nothing to disclose.

Dr. Pickering: Thank you so much, Risha.

Emma Hoo?

Member Hoo: Good morning. Emma Hoo with the Purchaser Business Group on Health. Nothing to disclose.

Dr. Pickering: Great. Thank you so much, Emma. I believe, Emily, we haven't received the MSDOI from you. It's that measure-specific one. The team will send an email to you right now. If we could just get that from you quickly --

Member Hoo: Sure.

Dr. Pickering: -- we can get through the proceeding. Thank you so much.

Okay. Sean Hopkins?

Member Hopkins: Yes, Sean Hopkins with the New Jersey Hospital Association. Nothing to disclose.

Dr. Pickering: Great. Thank you so much, Sean.

Jonathan Jaffrey? Jonathan Jaffrey?

Okay. Dinesh Kalra? Dinesh Kalra?

Okay. Suman Majumdar?

It looks like he's inactive. Just making sure.

Suman Majumdar?

Okay. Alefiyah Mesiwala? Alefiyah Mesiwala?

Okay. Pamela Roberts?

Member Roberts: Pam Roberts from Cedars-Sinai. Nothing to disclose.

Dr. Pickering: Thank you so much, Pam.

Mahil Senathirajah? Mahil, are you on?

Okay. Matthew Titmuss? Matthew Titmuss?

And Sophia Tripoli, we received a notification from her just earlier this week. We weren't able to update the slides, but she is going to be resigning from the Standing Committee. But just double checking here.

Sofia Tripoli?

Yes. So she's no longer going to be on the Standing Committee. We received notification earlier in the week.

Okay. Danny Van Leeuwen?

Member Van Leeuwen: Hi. Danny Van Leeuwen, Health Hats. I have been on an Acumen technical expert panel, but nothing related to any of these measures.

Dr. Pickering: Great. Thank you so much, Danny.

Thank you all again for your attendance and presence today. I'll just go back to the names that I didn't hear from. If you're on the call or joined late or had trouble getting off mute, please go ahead, and just now we're going through disclosures of interest.

Robert Bailey?

Lindsay Erickson?

Jonathan Jaffrey?

Dinesh Kalra?

Suman is inactive.

Alefiyah Mesiwala?

Mahil, if you're on.

Or Matthew Titmuss?

Okay. All right. Well, thank you all so much. I'd like to let you know if you do believe you have potential conflict at any time in the meeting as topics are discussed, please speak up. You may do so in real-time during the meeting, or you can send a message via chat to the chairs or anyone on the NQF staff.

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

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

Okay, thank you. As a reminder, NQF is a non-

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

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

With that, I will turn it back over to the team. LeeAnn, over to you.

Ms. White: Thank you everyone for joining the call today. I want to, in the spirit of transparency, let you know that we have ten Standing Committee members present today. We needed nine Standing Committee members to host a call, so thank you all for being here today. We greatly appreciate your participation and attendance.

We do not have Standing Committee members to meet the quorum for voting, which is 12 voting Standing Committee members. So our team is prepared. We will send out a Survey Monkey offline voting link to those Standing Committee members that are currently present on today's call. You are welcome to vote offline using that link.

To let you know the process following the meeting today, we will send out the recording of the meeting and a video to all Standing Committee members with the Survey Monkey link to those that are not able to attend today with a deadline for voting to be completed by. That will follow today's call. But again, thank you so much for being here today. We were glad we're able to review these measures.

This is essential. If you need to step away, please make sure you let us know and then when you return so that we can maintain high on our attendance numbers. If we drop below nine, then we will not be able to hold the call.

Thank you again, and I will hand it over to Isaac Sakyi, who will be reviewing our evaluating process and the voting process with the offline voting. Isaac?

Overview of Evaluation Process and Voting Process

Mr. Sakyi: Thank you, LeeAnn. I'll review the evaluation process that will be followed today. Our Standing Committee members act as a proxy for the NQF stakeholder membership. They evaluate each measure against each criterion and indicate the extent to which each criterion is met and the rationale for the rating.

They also respond to comments submitted during the public commenting period, make recommendations regarding endorsement to NQF membership and also oversee the portfolio of measures.

Next slide. To go over some ground rules, we'd like to emphasize that this is a shared space and there's no rank in the room. We encourage you to remain engaged in the discussion without distractions and hope you're prepared and have already reviewed the measures.

Please base your evaluation and recommendations on the measure evaluation criteria and guidance. Keep your comments concise and focused. Be cognizant of others and make space for others to contribute to the conversation.

Next slide. In terms of how the discussion will proceed, we'll start with an introduction of the measure by the measure developer. The lead discussant will then briefly explain the information provided by the developer on each criterion.

This will then be followed by a brief summary of the pre-evaluation comments from the Standing Committee, which will emphasize areas of concern or differences of opinion.

The lead discussants will also note preliminary ratings by NQF staff, which is intended to be used a

guide to facilitate the discussion. Developers will be available to respond to questions from the Standing Committee.

Afterwards, the full Standing Committee will discuss, vote on the criterion if needed, and move onto the next criterion. In this case, the voting will take place offline.

The following is a list of our endorsement criteria. Five areas are outlined here. Namely, importance to measure and report, which includes evidence and performance gap. Scientific acceptability, which includes reliability and validity. Please note that the first two bullet points are must-pass criteria.

We also have feasibility, usability and use, and related or competing measures. The Use subcriterion is a must-pass criterion for maintenance measures.

The next point of discussion is the comparison to related or competing measures, which is a discussion and does not require a vote. A discussion only takes place if the measure is recommended for endorsement.

Next slide. Here's a list of criteria the measures are evaluated and voted on.

Next slide. If a measure fails on one of the must-pass criterion, there's no further discussion or voting on the subsequent criterion for that measure. The Committee's discussion will move onto the next measure if applicable.

If consensus is not reached in a criterion, the discussion will continue to the next criterion, but there ultimately won't be a vote on the overall suitability for endorsement. This is of course the process should be doing a live vote with Poll Everywhere.

Next slide. As far as achieving consensus goes, a quorum is 66 percent of active Standing Committee members. And as mentioned earlier, that is 12 out of

18 active Standing Committee members for this committee.

We need greater than 60 percent of yes votes to pass a criterion or recommended measure for endorsement. Yes votes are the total of high and moderate votes. Forty to sixty percent of committee members voting yes will be consensus not reached, and less than 40 percent voting yes means the criterion does not pass or the measure is not recommended depending on what we're voting on.

Measures for consensus is not reached, we'll move forward to the public and NQF member comment period, and the Standing Committee will revote during the post-comment web meeting. If a measure is not recommended, it will also move onto the public and NQF member comment period.

The Committee will not revote on the measure during the post-comment meeting unless the Standing Committee decides to reconsider based on submitted comments or a reconsideration request is submitted by the developer.

As mentioned before, please let us know if you need to step out of the meeting. We need nine Standing Committee members to continue the discussion. At this moment, we have ten Standing Committee on the call.

Since we don't have quorum for online voting, an email will be sent out containing a link to the Survey Monkey. The Standing Committee members will be given 48 hours upon receiving the survey and a transcript of the meeting.

The Standing Committee members are on the call can follow the discussion as we move along and also submit their votes. When the meeting is over and we have access to our transcripts, that will be sent to the rest of the Standing Committee members to also cast their votes.

Next slide. At this moment, I'd like to pause, see if

there are any questions.

Hearing none, I'll turn it over to LeeAnn.

Measures Under Review

Ms. White: Thank you, Isaac. Please let us know if you have any questions along the way today. We are happy to assist. I do want to announce that a Survey Monkey offline voting link was sent to the Standing Committee members present on today's call at approximately 10:20 a.m. Eastern Time. Please let us know if you do not see the Survey Monkey link in your email inbox and we can resend that for you.

I'm going to hand it over -- actually, we're not doing a voting test with the live voting. We did not reach quorum, so next slide, please.

Okay, I will go through the measures that we will be reviewing today.

Next slide, please. The Cost and Efficiency project team received three new measures for the spring 2022 cost and efficiency cycle. Those measures are listed here. 3623, Elective Primary Hip Arthroplasty Measure. 3625, Non-Emergent Coronary Artery Bypass Graft Measure. And 3626, Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure.

Next slide, please. I'd like to take a moment to review the Scientific Methods Panel. The Scientific Methods Panel is a group of researchers, experts and methodologists in healthcare, quality and quality measurements.

The Panel reviews complex measures and provides comments and concerns to the developer. The developer then has the opportunity to further clarify and update their measure submission prior to the Standing Committee evaluation.

Next slide, please. The SMP evaluated the scientific acceptability of all three measures that we'll be

reviewing on our call today, and all three measures passed the SMP evaluation review.

Next slide, please. With that, we'll now begin the review of our first measure. Our co-chairs will start to begin by introducing the measure. The developer will then have an opportunity to provide a three to five minute overview of their measure.

Our lead discussants will introduce the criterion and highlight their main take-aways. Our supporting discussants will respond to the lead discussant and add their insights.

During the criterion discussion, the co-chairs and staff will collect all questions for the developer whether they're verbal or in the chat. We'll collect those during the discussion.

Once the initial discussion on the criterion is complete, the co-chairs will then allow the period of time for the developers to respond to those questions and clarify any information.

Once the Standing Committee has completed its discussion, a vote will be taken on the discussed criterion.

I will pause for a moment to see if anyone has any questions before we start our measure review.

Okay, I'm hearing none. I also want to pause to see if our measure developer is on the line today.

Do we have a member of the developer team from Acumen on the call today?

Mr. Bounds: Yes, we have a couple. This is Sam Bounds. We have Joyce, who's going to be measure intros. She's here. Ken and Pickering.

3623 Elective Primary Hip Arthroplasty Measure
(CMS/Acumen, LLC)

Dr. Pickering: Wonderful. Thank you so much. Good morning. Thank you for joining us.

Next slide, please. I will hand the baton over to our co-chair, Kristine Martin Anderson, to lead us in the discussion of Measure 3623, Elective Primary Hip Arthroplasty Measure.

Kristine?

Chair Martin Anderson: Thank you, LeeAnn.

As you can see on this slide, we are starting with 3623. Acumen will introduce this particular measure. I think it's probably best that I'll just read the screen to you because they're going to tell you the relevant elements of the measure. And then after Acumen introduces the measure, we will be turning it over to Amy Chin, who's going to be our lead discussant to take us through the criteria.

Joyce, let me turn it over to you.

Ms. White: Joyce, if you're speaking, we can't hear you. I believe you're on mute.

Chair Martin Anderson: We can't hear you still, Joyce, but it looks like you're trying to work it out.

Dr. Pickering: We will have a member to reach out to Joyce to help troubleshoot the audio. Is another member of the developer team ready to provide an introduction to the measure?

Mr. Bounds: Yes, this is Sam from the developing team. I'm pulling it up right now.

Thank you for your patience as we worked on our audio and kind of quick handoff. Let me do a brief introduction.

Good morning, and thanks for the chance to provide an introduction. As you know, it takes a lot to develop and maintain these measures. I want to introduce a few members on the line that are going to help today.

Joyce, as we get her audio connected, she's the project manager. We also have Dr. Heather Litvinoff, our clinician lead. Myself and Ken Tran are technical

leads. And Sri Nagavarapu is our statistical advisor and project director.

I'll start by providing a summary of how the measure is constructed then briefly describe the development process and how the measure is being used in MIPS.

Our first measure that we're starting with today is the elective primary hip arthroplasty cost measure. It evaluates a clinician's risk-adjusted cost to Medicare for care related to this surgical procedure.

When we talk about cost for this, and the other measures today really, it refers to the allowed amounts in Medicare claims data, which includes both Medicare trusted fund payments and any applicable beneficiary deductible and co-insurance amounts.

We also use payment standardized cost, which assigns a comparable amount for the same services regardless of geographic area and other factors that aren't related to healthcare delivery choices such as a graduate medical education payment.

Care for hip arthroplasty is assessed through episodes of care which are initiated or triggered by a CPT/HCPCS procedure code for hip replacement in the outpatient or inpatient setting.

This trigger procedure is used to identify the episode window or the period in which care is assessed. The episode window starts 30 days prior to this trigger and ends 90 days after.

During the episode window, the measure only includes costs that are clinically related services that can be influenced by the clinician's care decisions. Examples include pre-operative workup imaging, wound care, complications, routine follow-up and other consequences of care.

The trigger procedure also is used for attribution. So simply enough, the individual clinician billing the procedure, the surgeon is attributed the episode.

Episodes attributed to individual clinicians are rolled up to the group practice-level, or when look at the codes, that's the individual being identified by a TIN/NPI and then being rolled up to their clinician group identified by a TIN.

This is rolling up to that TIN clinician group-level is actually how the vast majority of clinicians currently participate in MIPS, which this cost measure is intended for.

The measure takes into account the patient case mix through the use of risk adjustment. So the risk adjustment model includes many variables including those from the CMS-HCC model, which is used in Part C.

Reasons for enrollment, interactions terms of those HCCs, age bracket factors and other factors affecting cost for this specific procedure such as a history of spinal disorders.

So while we start with a base risk adjustment model like the HCC to get a lot of bandwidth and clinical beneficiary case mix, we also customize it to the cost measure itself when working with our workgroup.

And then finally to create a score, we first calculate the observed cost or all the costs that occurred during the window over the expected costs. The expected cost is as predicted through our risk adjustment model for each episode that's attributed to a clinician.

Ultimately, the clinician's score is just an average of those observed to expected ratios across the episode sample that's attributed to them, which we can then translate to dollar amounts using some mean observed cost scalar.

Ultimately, clinicians are being compared to their peers with a lower score representing better performance. I'm still going to go through development process and important use in alignment with quality. I appreciate this opportunity.

I do want to flag that measure construction section shares a lot of similarity with the other two measures that we're going to look at today. So taking a little time on this first one will probably help us catch time back up later on today.

I do want to go over the development process of these cost measures and the specificity that external clinicians put into creating these measures.

This measure was selected for development in 2018. We used criteria such as frequency and its high cost procedure for the Medicare population and also aligns with the NQF endorsement knee arthroplasty cost measure.

The specifications were built through a robust development process that incorporated input from 44 clinical experts, a technical expert panel with 19 members, and eight individuals representing the patient and caregiving perspective.

We met with the development panels five times over an approximately 12-month period and iteratively conducted empirical testing as we continue to develop measures and provide feedback to the workgroup and listen to their responses and have good discussion.

We also held a national field testing period to gather input from a broader range of clinicians who would be attributed the measure. So this is beta testing.

We produced over 8,000 field test reports containing performance results, how you score on the measure, how the measure is constructed, but also all this great feedback information on where your costs are coming from, what your patient case mix looks like, et cetera, et cetera.

Lastly, I want to talk about the important use and alignment with quality. After development and pre-rulemaking of the rulemaking process, the measure was added to the MIPS cost performance category for 2020 and submitted the spring 2021 cycle. And the

SMP, or the Scientific Methods Panel, rated it high on reliability and moderate for validity.

In MIPS, clinicians receive a cost performance category score which is the average of their performance on each of the cost measures for which they meet the case minimum. So this would go into the MIPS cost category and be a collection of measures that a clinician may receive based on their care patterns.

In 2022, the cost performance categories weighted at 30 percent of the MIPS final score by statute, which combines scores from other categories for quality measures, improvement activities, and the use of the electronic health records. So the intent of this cost measure is to be used in combination with the quality and improvement activities in the MIPS program.

The hip arthroplasty measure is included in a MIPS value pathway or a MVP for improving care for lower extremity joint repair, which will be available for reporting in 2023.

This means that clinicians reporting for this measure value pathway, this MVP, would be evaluated on this cost measure plus hip replacement-specific quality measures like complication rates, functional status and evaluation of cardiovascular risk factors prior to surgery.

So that's kind of a new cool idea in MIPS in which they're really starting to pair these quality and cost measures together and kind of specialty-focused evaluation MVPs.

While the cost performance category of MIPS was reweighted for 2020 and 2021 due to the impact of COVID-19 public health emergency, there will be beneficiary-level reports available for 2021.

Thank you again for the chance to introduce the measure. Since the review of this measure was deferred by NQF from last year to this cycle, we've

conducted some additional testing since then, since the original submission.

This includes further testing on social risk factors, which we're happy to talk through to provide the Committee with all the information they need in their discussion today. Thanks again.

Chair Martin Anderson: Thank you for that overview.

Now, I'm going to turn it over to Amy Chin, who will be supported by Pam Roberts to go through the criteria and get the feedback from the committee.

Member Chin: Hi, everyone.

Do I need to introduce the measure again? I know it's in the script, but --

Chair Martin Anderson: I think you do not since it was just introduced.

Member Chin: Okay, great.

Going through the first area or evaluation criteria importance to measure and report. So Part 1A, we look at the evidence provided. Based on the evidence -- the evidence mainly covers that this is a high-prevalence condition, total hip arthroplasties in the U.S. with an increasing number of Medicare-age patients receiving total hip arthroplasties through 2030.

The developer also notes that there is higher variation in treatment options for total hip arthroplasties, so that represents an opportunity to improve cost efficiency and quality of care, especially in the area of appropriate use of institutional post-acute care. So that's institution care versus care at home or outpatient therapy. They also point towards trying to increase the use of optical surgical techniques.

So based on the pre-evaluation, we also -- let's see. I don't see -- oh, the staff preliminary rating on this is moderate for this area, and I agree with that

rating. I recommend that rating.

Pam, I don't know if you have thoughts?

Member Roberts: No, I think actually you covered it well. I agree with you with the moderate rating.

Member Chin: Can we open it up to discussion at this point, or do we just move forward?

Chair Martin Anderson: Yes.

Member Chin: Okay. Any thoughts from the rest of the committee on this area?

Member Gidwani: Hi, this is Risha Gidwani. I have a question for the developer.

I was looking at the supplemental information, and the national summary data report from 2018 showed that between the 25th to 75th percentile, there was a $3,000 cost difference.

Let me get off video you can see me. Hi.

And then from the 10th to the 90th percentile, there was a 50% cost difference. I was not able to tell from the documents provided how many patients are actually receiving this procedure a year.

Can you provide that so we can get some back-of-the-envelope math as to the extent of cost savings that could occur if everyone moved, say, from the 90th to the 10th percentile?

Ms. Lam: Hi, this is Joyce. Can you hear me now?

Chair Martin Anderson: Yes.

Member Chin: Yes.

Ms. Lam: Okay, great. Sorry about that. Yes, let me just pull up some of those numbers. We do have the number of providers in each of the deciles, but I'll look up to see if we have the number of episode and things.

Ms. Lam: So for hip arthroplasty, there's about 115,000 beneficiaries overall. So we did rerun the measures on some more updated data. So for fiscal year 2021, there's a total of 125,468 episodes.

Chair Martin Anderson: Danny, I see you have your hand up?

Member Van Leeuwen: Yes, thank you.

So I'm not a statistician, but I'm looking at this opportunity, 1B opportunity for improvement. They're talking about that the data indicated a mean score of 1.03 with a standard deviation of 0.12 and an interquartile of 0.15.

I know what all those words mean, but I don't know what this means altogether. I don't know how to place this is in -- is this enough variation to be worthwhile?

Chair Martin Anderson: I think that's the same question that Risha was just going after, right. So if it's about a 15 percent difference, the question is given how many dollars that is and then how many cases there are, is there enough variability to make it important.

I don't know, Risha, if you did the math already for what percent would be in --

Member Gidwani: Yes. If everyone moved from the 75th to the 25th percentile of cost, then it'd be a cost savings of $375 million to Medicare.

Chair Martin Anderson: Does that make sense, Danny?

Member Van Leeuwen: Yes, that makes sense, but they're not going to. That's great to say that. That's the outside end of a continuum of possibility. But it doesn't -- anyway, I can move on.

Member Chin: I --

Chair Martin Anderson: Who's trying to speak?

Member Chin: Amy.

Chair Martin Anderson: Oh, I heard Danny.

Let me just raise another element in the same vein. I think later in the criteria, Danny, we deal with, is it improvable? Right, in terms of can you use this measure to actually achieve the improvement.

But I had a question around the correlation with other measures. I see you chose another cost measure. But for the importance criteria, et cetera, it specifically asks about correlation with quality measures.

I know there are other quality measures the developer introduced in the introduction that the MIPS program will also use for this particular measure.

Have you already looked at how this measure performs relative to those? In other words, is it even feasible to say that lower cost is better? Do we have any sense of that?

Ms. Lam: Thanks for the question. We do have some information about the analysis we've done exploring the relationship with quality. You might see some of that in the empirical validity section. But before I talk about the analyses, there is a wealth of literature about lower extremity joint replacement.

So one of the reasons that we were interested in this measure is the comprehensive joint replacement model, a CMS Innovation Center model, also abbreviated to CJR. They have an annual report which found statistically significant decreases in payments mostly due to reductions in institutional PAC use. The model focuses on knee and hip arthroplasty.

Some of the particular findings were that fewer patients discharged and dismissed but fewer days there and more patients were discharged to home health. The really encouraging this is that the quality of care was maintained or improved as measured

through some of the quality metrics available in CJR such as unplanned readmission rate, ED use and mortality.

They also looked at patient satisfaction. And so the evaluation found that few patients had similar satisfactions recovery and care than non-CJR patients. These are some of the reasons from CJR that point to the opportunities for cost improvement while maintaining quality.

For the relationship with other MIPS measures, that is something that we've looked into. The reason it is challenging to give you analyses is that the MIPS quality performance category, it's all voluntary.

So MIPS participants pick usually six quality measure including one outcome or high priority measure. So there are challenges around small sample sizes, the potential bias from selection as well as data completeness where currently participants only have to report 70 percent of data for eligible beneficiaries in the denominator.

Having said that, we have looked at the correlations with the quality measures that were mentioned in the introduction which do focus on hip replacement-specific aspects of care such as functional status assessment.

Unfortunately, the number of TIN and TIN/NPIs who report that measure plus who are attributed under our cost measure, it's a very low count. So for instance, the Measure 376 functional status assessment for total hip replacement, we only have 92 pairs of TINs and 444 TIN/NPIs. And none of the results that we looked at were statistically significant.

There is the risk standardized complication rates measure in MIPS. That was only implemented in 2021. That is a claims-based measure, so it doesn't suffer from the same problems with selection bias or people choosing report it because it's automatically calculated.

That is a measure we've been keeping an eye on because of the close connection with our measure. Unfortunately, we haven't had a chance to do analyses with it because the data for 2021 for MIPS quality measures has only been finalized at very end of June.

That's something that we could explore more for the post-comment evaluation meeting. But knowing that we weren't able to get that data in time, we did do an analysis where we reconstructed the numerator from that hip arthroplasty MIPS measure.

We tried to provide some information where we took the definition of the numerator from that measure which counts complications as AMI, pneumonia, sepsis, surgical site infection, pulmonary embolism and certain other serious complications on an inpatient claim.

So we calculated this complication rate for each attributed provider. Again, it's just taking the numerator definition from that measure and we calculated correlations between that complication rate and the cost measure score.

As we expected, providers who tend to do well on the cost measure also tend to have lower complication rates. The correlation that we sort of -- Pearson correlation is 0.27, so a medium correlation of both the TIN and TIN/NIP levels.

Chair Martin Anderson: Thank you.

Amy or anyone else have any more questions?

Member Chin: I think Risha has her hand up right now.

Oh, Risha. You're muted.

Member Gidwani: Sorry, thank you.

I think your last sentence was, I think, a very important one. Maybe we can unpack that a little bit more. You said that there was a correlation of 0.27

amongst providers that have lower costs and had also lower complication rate.

It's a pretty low correlation, 0.27. Do you have any information that you can present to this committee as to, let's say, the difference in providers at the 25th and 75th percentile or 90th and 10th percentile as to the rates of, let's say, hospital readmission or need for revision surgeries?

Dr. Pickering: Sorry, this is Matt from NQF. Apologies and sorry, Risha. Great question. I just want to make sure that we're not getting too far into validity testing and correlations because that's going to be coming up next.

I understand there's definitely some considerations and importance of the measure as it's sort of correlating to quality indicators. However, we don't want to go too far into the testing and evaluation of the analyses just because that's reserved for the validity testing.

Member Gidwani: Sure, okay.

Dr. Pickering: Just a reminder, we want to kind of keep to is there a high impact or high resource area here, and is there really variation that we see with this measure.

Just reminding the folks. Not to cut everyone off because this is great dialogue, but validity testing is where we can get into some of those discussions.

Member Gidwani: Okay, we can punt on that question, then. I do have another question, which is more global. And Matt, you can tell me if I need to save that for another time, but I think it might work here as well.

That's just in general, do physicians have the ability to move the needle on this? So when they looked at the services that were included, it really seems like some ways to get some good cost-savings here are to do things like move the surgery from inpatient

setting to an ambulatory care surgery center, or to be judicious about where you are asking your patient to get their imaging done so that you're going to a lower cost facility.

But the question I have is whether physicians really have the ability to change this. Over 50 percent of physicians in this country are now in a practice that is owned by a hospital or a healthcare system. Those hospitals and healthcare systems oftentimes will set up situations such that they are able to get the highest reimbursement, right.

They have the incentive to make the surgery happen in an inpatient setting if they're going to get more reimbursement for it. They have the incentive to have the MRI happen on their campus so that they can get more reimbursement.

That's not to say that we shouldn't try to ensure that services are provided at lower costs if they're of equivalent quality, but the question is whether the actor in this situation, the recipient of this measure, is the one that has the ability to move the needle on it.

So just sort of giving this large buyout of physician practices that just happened in the last years and has accelerated in the last couple of years, I welcome hearing from Acumen as to whether you feel as though --

Chair Martin Anderson: Risha, I'm going to hold that one as well because that's really usability. The question is, is there a gap? That's question one. Can it be closed is -- is there a valid measurement is reliability, then can it be closed as usability. We captured what you said, and we will hold that as well for that part.

Member Gidwani: Okay.

Chair Martin Anderson: Any other questions on importance or availability for improvement?

Since we can't vote, you can put it on your Survey Monkey, but I will turn it back to Amy to get us into what we're interested to talk about, which is scientific acceptability.

Member Chin: Okay. So this measure was reviewed by the Scientific Methods Panel. Looking at reliability in terms of specifications and reliability testing.

In terms of specifications, the measure is well-defined. It's a claims-based measure. A lot of that will be very consistent and standardized in terms of how it will be measured and whether the measurement when it's repeated will be the same as previously.

In terms of reliability, the developer conducted a signal-to-noise analysis and split-sample analysis. They did that the TIN and the TIN/NPI-level.

In general, the Scientific Methods Panel found everything reliable and passed the measure with a high rating for reliability, so that's seven members voted high, one medium and zero low.

I agree with the Scientific Methods Panel that the approaches were appropriate and the testing results indicate high measure reliability.

Chair Martin Anderson: Can you summarize any of the comments that came from the committee? Is there anything worth noting that we got in the pre-evaluation comments?

Member Chin: There was one. Let's see.

The primary comments were really around the measure specifications in terms of conceptualizing it so that the attribution to a main assistant or clinician, how do you decide what is the main or assistant clinician if both are found in the claims.

And then in terms of assigning costs to an episode and calculating total observed episode cost, how were related and unrelated services determined for the episode.

Chair Martin Anderson: Okay. Why don't we start with reliability and see if Pam or anyone else on the committee has a comment on reliability that's related to specifications and testing.

Member Roberts: I don't have anything else to add.

Chair Martin Anderson: Thank you. The committee find the specifications adequate. Any questions on those?

Okay, and then reliability testing. Any questions on how the reliability was tested or the results?

Member Gidwani: I have a question about the assistant versus main clinician. Can you tell me if there's both an assistant and a main provider how the attribution works for them?

Mr. Bounds: Is this for the developer --

(Simultaneous speaking.)

Mr. Bounds: -- jump in?

Member Gidwani: Yes.

Mr. Bounds: All right, good.

Yes, we can use modifier codes to identify the main and the assistant surgeon. They'll actually bill a Part B Physician/Supplier claim with the surgery in that event, and both will be attributed the same episode. There's joint responsibility from each clinician in terms of measurement in this cost measure.

Member Gidwani: So the entirety of the episode cost is given to both clinicians?

Mr. Bounds: That's correct.

Member Gidwani: I have to say I'm not a clinician, so I don't really know how this works. But if there are any physicians on this panel, I'd welcome hearing whether the assistant surgeon has the same ability to influence cost as the primary.

So for example, the discharge planning, discharging to home with home health versus discharging to a SNF. Is that decision made jointly by the assistant physician and the main physician, or is that the primary surgeon who has that responsibility?

Member Van Leeuwen: Well, this is Danny. I can speak as a nurse. I can say often neither. It's often the social worker, the nurse who is active in follow-up more than the surgeons. Either the primary or the secondary.

Member Gidwani: Thanks, Danny. Does that also apply to making the decision about where the patient goes post-discharge?

Member Van Leeuwen: In my limited experience, yes.

Chair Jhamnani: So maybe I can give some clinical insights. This does not apply to THA to TKA, but in clinical practices in general. A discharge disposition or planning is determined in several criteria. Many of the clinical elements of the patient, how the patient was doing, mainly for TKA/THA, how the mobility is assuming the surgery went well, there were no complications, what physical therapy, occupational therapy says, there's social issues.

It's a wide complex of factors which determine where and when the patient goes. So it's a composite. So part is determined by the attending physician, who's either the main surgeon or the assistant surgeon. It depends on the hospital, it depends on the care providing who makes that decision and that the other factors that I just mentioned.

It's hard to give a clear-cut answer as to who makes that decision because it's usually a team that makes the decision.

Chair Martin Anderson: Pam, you wanted to join in?

You seem to be on mute, Pam.

Member Roberts: Sorry. You'd think I'd learn by now. A lot of times it's led by the hospitalists, and as was just mentioned, it's really a team decision a lot of time at the point of the developer determining where they go in things. It's really an effort with an entire institution. The surgeon or the assistant surgeon may not have much to do if everything goes well.

Chair Martin Anderson: Okay, Joyce. Do you have something to say that's short on this?

Ms. Lam: Yes, just really quickly. The assistant is paid a portion and is part of the surgery. The way that we do attribution really recognizes this idea that it is a team, so we hold both the main and assistant clinician responsible for this care so they are incentivized to coordinate on the care.

The other point I'll just mention in terms of how we built the specifications is the input that we got from the person and caregiver perspective where we heard from PFPs, patient and family partners, where the surgeons were leading the care team and they received discharge planning care from the surgeon and nurses.

Really, it reflects this idea that everyone is part of the team. And by holding the clinicians who perform the surgery, both main and assistant, accountable this reflects that input that we got from the PFP perspective.

Chair Martin Anderson: Okay, I'm going to pause for a moment here if anyone's got their Survey Monkey open and wants to reflect their viewpoint on reliability. And then remind everyone reliability is specifications are well-defined and been precisely specified and that the testing is repeatable, right, that when you apply this same method over and over again, you're going to get similar results and there was a whole set of testing that you all read about for how that was done.

And then we're going to move to validity, which is whether or not the measure specifications are

consistent with the measure intent and captures the population and the testing that was done that the measure is correct or correctly reflects the cost of care or resources and also exclusions, risk adjustment, meaningful differences, et cetera. It's much more robust.

Amy, do you want to say anything specifically about validity before we open it up?

Member Chin: Yes, sure. So just want to share the Scientific Methods Panel findings on validity. They voted one high, five medium, two low. Just a quick overview of the methods used for validity testing, there was testing with both empirical validity testing and then with a TEP to assess face validity.

One point, the Scientific Methods Panel brought up is that they only tested one measure in terms of the empirical testing, and that was a hospital-based measure Medicare spending per beneficiary.

Chair Martin Anderson: Committee questions or comments?

I'll start with one. I think we keep running back into this social risk factor issue measure after measure. Maybe, Matt, you'll tell me soon that the committee that was going to meet at NQF solved this issue.

I read all of your rationale for why you didn't think social risk factors should be in the risk adjustment itself. Did you stratify the results so you could see whether or not the actual results differed by some of the social risk factors so that we could have a sense of whether or not there actually are disparities in care?

That's for the developer.

Mr. Nagavarapu: This is Sri from Acumen. Can you all hear me okay?

Chair Martin Anderson: Yes.

Member Chin: Yes.

Mr. Nagavarapu: Okay, great. Thanks for the question. This is an area we're really keen on exploring and make sure that the measures are performing as expected. So we kind of go through a standard battery of testing for all the measures.

For some of them, we end up adjusting for social risk factors based on the empirical results. For some, the empirical results suggests it could cause more harm than good to adjust for social risk factors.

Part of that testing is to stratify the measure results along both individual and sort of provider-level dimensions as you're suggesting. For the individual-level dimension for instance, it's stratification of risk-adjusted episode cost by whether a beneficiary is dual eligible or not.

And for the provider-level, just to give you the analogue for duals, it would be a stratification by the chair of a provider's episodes that are dual-eligible or not.

And something that we see as an interesting pattern that makes us hesitant to risk adjust for measures where that pattern shows up is that often for a set of measures like this, that stratification is useful and sort of depicting the differences between risk-adjusted cost between dual episodes and non-dual episodes is smaller and sometimes swamped by the difference across providers with high dual shares versus low dual shares.

So the basic pattern you see that Joyce and Sam may talk through in more detail because we have detailed empirical results on this, but real quickly the basic pattern you see is that in general as you go to providers with higher dual shares, the risk-adjusted cost for duals either increases or remains stable.

But the really interesting point that makes us very wary about how to approach us with a kind of a do-no-harm approach is that the risk-adjusted cost for non-dual episodes also increases as you go through providers that have a higher dual share.

And so the concern is that providers with a higher dual share are systematically performing worse both for duals and non-duals. I think the stratification that you're alluding to has sort of allowed us to see that.

And so an implication or risk adjusting for dual status, for instance, would be that you might remove some of the difference in performance that's actually due to a provider-level effect rather than individual-level effect because you do see that same sort of pattern of increasing risk-adjusted cost for both non-duals and duals.

Chair Martin Anderson: Thank you, that's very helpful.

Danny.

Member Van Leeuwen: Yes, I am interested in that the exclusion criteria of less than ten episode and its relation to disparities. So I'm wondering if any thought has been given to the disparities in the proportion of patients who go to clinicians that have done less than ten visits per period or episodes per period than those that have done more than ten episodes, and the relationship of cost on that.

Mr. Nagavarapu: Thanks for that question. We haven't looked specifically at that. I think that is an interesting point that sort of the size of episode count and the experience of a surgeon could vary according to social risk factors.

The main reason we've stayed away from the specifics of the type of analysis you've mentioned is in order to maintain reliability, we kind of focused all the analyses on those with episode counts above ten. But if that would ever change or if there's specific request from the committee, we could certainly look at that kind of pattern that you're talking about.

Member Van Leeuwen: Thank you. As a patient representative, certainly a really important factor for me is how many cases have the person gone to.

It's just interesting the tension between trying to do reliability and meaningfulness and then what does it mean to me as a patient these cost measures that this whole world is excluded of less than ten. I don't even know what proportion that is.

Anyway, it's just an interesting -- thank you for you --

(Simultaneous speaking.)

Chair Martin Anderson: And I do believe that I also read in this measure that there was not a relationship between results and number of cases, right? That struck me the same way, Danny, but it said it did not vary by region nor by number of cases.

Ms. Lam: Yes, that's right. So, in the CR (phonetic) in the testing form, we provided a range of stratifications by different characteristics like geographic regions. We did find that the scores were stable across --

(Pause.)

Chair Martin Anderson: Which might get to the systems of care issue that may trump a little bit on cost measures.

Okay, any other -- oh, I see Risha.

Member Gidwani: Hi, one last question on the risk adjustment. Was there any risk adjustment that was done and not reported related to rural versus urban status, or maybe higher resources versus lower resource areas.

If providers don't have options to send someone to home health because home health isn't available in their geographic area, I'm sort of wondering about the implications that that would have for scoring. Did you guys do any investigations into geographic area or rurality.

Ms. Lam: Yes, this is in Table 2B42 of the testing appendix where we looked at the provider

characteristics of investors (phonetic) rule. We found, like with all the results, generally on that tab which looked at provider characteristics. But it was very similar across them. So the mean score for TIN in urban, 1.03, and for rural, 1.02. It's showing rural providers on average tend to do a little bit better. And at the TIN/NPI-level, it was 1.00 for urban and 0.99 for rural.

Member Gidwani: Thank you. Can I ask you to direct to me where that is, or maybe NQF. Can you guys let me know where I can find that information?

Was that provided in the measure specifications we were given?

(Simultaneous speaking.)

Dr. Pickering: Yes, we're looking -- sorry, Joyce, were you going to say something?

Ms. Lam: No, go ahead.

Dr. Pickering: Joyce, can you say that table number again?

Ms. Lam: Oh, sure. It's in the testing appendix. It's a workbook, and it's got a number of tabs. It's the final tab, which is called Table 2B42, score by provider. That's in the testing appendix that was part of the submission.

Dr. Pickering: Risha, we'll follow up with that. If there's any other questions for the developer?

Member Gidwani: Yes, I don't think -- if that's an Excel workbook, we were not provided that as the committee members.

Chair Martin Anderson: Okay. Are there any other questions on validity?

Chair Jhamnani: I have one. First question, and this is a question which pertains to these measures. One of the struggles I have is assessing face validity as we all talk and serve in other domains.

When you don't have a lot of people, when the sample size is low, evaluating any results, what happens out of that test is questionable. For example if you have a sample size of 11 people who are answering questions on face validity and when your total number of surgeons based on your TIN/NPIs are thousands, whatever they say, those 11 members were members of the TEP. How much do I trust them?

It boils down to the question of what should be an ideal size of a TEP evaluating for face validity. Because face validity by itself is not a very robust analysis of validity, even though it's one of the measures that NQF requires.

I have some problems trusting that even though there was good agreement on the 11, but are those 11 a good sample size to be accurately representative of the total number on the orthopedic surgeons in this country. I'm not sure I would trust 11 people.

I would rather like to see input from a larger TEP, maybe more than 10, it may be 11. I don't know if the helper (phonetic) have answers to that, but that was the process that was undertaken. I think I have some concerns about that. I don't need an answer on that yet.

The other question or issue I had was the empirical validity testing that was similar across all three measures. The way you did your empirical validity testing was to correlate with MSPB scores and you just ranged that from zero to ten, where zero score is a worse performance and the ten score is high performance.

If you look at correlations based on Table 4, which is in Page 60 of the handout that was provided by NQF, the correlation stuff. If you look at zero and the measure O:E was 1.12 and it was 5 to 10, which is a higher MSPB score, your O:E ratio mean was 1.

So the directionality seems to be appropriate, but

then is that good enough, number one. And number two, it seems like, yes, if you have a procedure where the cost is high, it usually tends to be as we talked to be a system performance. That tends to be at hospitals where other cost of care procedures are also higher.

It does make intuitive sense that even if you didn't provide me that data, I would have been able to come to that in an analysis by myself. So what I'd like to see other methods of empirical validity testing which broaden my horizon a bit more, and I don't see that.

The third question, and I'm lumping all my issues together so that we can save time and you can answer all the questions, is the issue of excluding death. This bothers me as a procedurist. I'm not an orthopedic surgeon; I'm a interventional cardiologist.

As all clinicians, I think we should be liable for our patients and the outcomes of our patients. For a procedure which looks at all costs of care or 90-day window, I think we're looking at readmissions, we're looking at where patients land up. Is it in the hospital, is it in inpatient rehab and stuff like that, complications.

But you're excluding the most important complication, which is death. I don't know how I feel about that. I understand that your rationale was that the costs are different, they're higher, but I did feel that they should have been included. It's something that providers and hospital systems and everyone should be accountable.

And if you do a bad job and the patient passes away, you should be accountable for that. You excluding that was something that I did not feel okay with. And none of these questions I don't think you can change. These are the way that you've done it.

But these are sort of the red flags I have, which is pretty much what led to my concerns with validity and not siding with the majority of the Scientific

Methods Panel, which was moderate, but in fact my analysis what is insufficient for me to gauge validity based on the testing you've provided.

Ms. Lam: Thanks for the questions. I'll go through each of them, but let me know if I didn't cover all of them.

So the first question that I heard was about face validity. So, absolutely. We have similar concerns, and so we've designed the process for developing episode-based measures that really get at this question of face validity and making sure that these measures are developed by clinicians for clinicians.

I think there's a lot of detail in the form about how much input that we got through this process. The first panel that we had convened was experts in musculoskeletal disease management, which had 29 member affiliated with 26 organizations and specialty societies, which included large medical systems, societies such as the American Association of Hip and Knee Surgeons, American Occupational Therapy Association and many others.

And so the role of this larger group was to identify which measures should be developed. So really in the measure prioritization looking at the impact opportunities for improvement. Through that committee, they recommended developing the hip arthroplasty cost measure.

After that initial process for prioritization and making sure we got that breadth of input for determining which measure would be the most impactful for this area of care, we convened a small web group as you mentioned.

This time, we intentionally aimed to have 15 members, which is what the CMS measures management systems blueprint recommends about the normal size for a clinical panel or a TEP. So these members were all focused on hip arthroplasty, so more specialized than the musculoskeletal disease management.

Again, we had a nomination period where societies, organizations, individuals could submit nominations for this panel. So making sure that we have the necessary expertise. Again, we had folks who were affiliated with the American Association of Hip and Knee Surgeons, American Academy of Orthopedic Surgeons as well as folks from across the care continuum. So American Academy of Physical Medicine and Rehabilitation, American Physical Therapy Association.

So a lot of input and expertise that we collected through that measure development process. As I mentioned in the introduction, we also held a national field testing period. This was to get that broader input that you had mentioned.

So we created a confidential feedback reports that were available to attributed clinicians. We created over 8,000 reports for this measure, and we received 67 responses to this. That was the responses for all the field testing, and there other measures and some comments applied across measures.

There's a really very in-depth iterative process that we've developed for making sure that as we built out these specifications, we are getting that clinical expertise.

Chair Jhamnani: Thank you, Joyce, but it still doesn't answer my question. The sample size at the end of the day is low. It's 15 clinicians, 11 clinicians. Although the representation was not, which is my point, and maybe CMS when they start reevaluating patients or groups for TEPs should probably consider a higher number of sample.

Being a clinician, I have 40, 50 cardiologists in my practice. If I ask a question to 40, 50 cardiologists, I'm going to get 40, 50 different answers. So to find a real truth in that question, I need a larger sample size. And I'm not sure 10, 15, 11, whatever number you choose there is good enough. That was my point. The process is robust. I'll give you that.

Mr. Nagavarapu: This is Sri from Acumen. I think the concern that you're pointing to, Dr. Jhamnani, and it's a concern with all measures developed according to the blueprint process. That's exactly what led us to the field testing that we undertook where we distributed reports to all attributed clinicians nationally.

There's 8,000 reports distributed and society has had a chance to view the actual operation and calculation of the measures, all of the mechanics, all of the specifications to be able to provide that broad base support that you're talking about.

Because exactly, we also felt that the import of any TEP needed to be complemented with something universal. This was sent out to the universe of clinicians to allow for a public comment period.

And we took public comments and addressed measure specification changes with the workgroup after that, too. So that group of 15 had the opportunity to receive input from a much wider set of clinicians for exactly the reason you're pointing to.

Chair Martin Anderson: Okay, thank you very much.

Risha, is your hand just up from your last question, or did you have anything else? I really think we should move on at this point. If anyone else out there has a question, you could put it in the chat so the developer will capture that for later. You can go ahead and make your notes in your Survey Monkey for validity if you wish. Let's move onto to feasibility.

Back to you, Amy.

Member Chin: Sure. In terms of feasibility, this is the extent to which the specifications including methodologic require that our data that are readily available or could be captured without due burden.

So this is a claims-based measure. So typically, highly feasible. The data is coded by someone else. I believe I would rate it as a high. The staff preliminary

rating was high as well.

So open it up to everyone else for comments on feasibility.

Chair Martin Anderson: Pam or anyone else?

Member Roberts: Amy is covering it well.

Chair Martin Anderson: Okay, on the usability and use.

Member Chin: Starting with use, this is the extent to which the audience could use the performance results that will be produced from this measure.

So the developer said that this is a measure that's going to be used in MIPs and used by clinicians and then also presumably consumers to help guide choosing a clinician. I guess part of the understanding of use would be from the clinician side and then also from the audience, which is the consumers as well.

From the clinician side in terms of use, I am slightly concerned about the mix of inpatient ambulatory surgery, outpatient and office settings mixed together.

I want to understand more about how that will be presented so that clinicians may understand that maybe they're using an inpatient setting more versus other settings.

The other part would be how is this tied with quality. Because I think currently, there's not that many quality metrics in these different settings and the regulations on performing these surgeries in each setting is slightly different.

If there's pressure for clinicians to move surgeries to, let's say, outside of the inpatient, what is the impact on quality and how would they assess that from the reports that are going to be provided.

Ms. Lam: So for the place of setting, we account for

the different settings through risk adjustment. So we included variables for inpatient stays and the specific DIG that it falls within. That's how we account for the differences in setting.

In terms of the information provided to clinicians, this is done through the MIPS feedback reports. We reached out to the contractors to get some more details about what is included in this.

They provide patient-level reports, which include a lot of detailed information including the post-episode identifiers plus a list of all the services received during an episode and the standardized cost for those services. And it's organized into various service categories.

For instance, there's information about post-acute costs for outpatient facilities. So there's a lot of granular information that providers can use to understand their performance and make practice changes.

Mr. Nagavarapu: For your question about unintended consequences, Dr. Chin, yes. As Joyce mentioned, because of the risk adjustment for inpatient versus outpatient, currently there's no penalty in the measure for doing a case inpatient. That was a choice that was made partly in consultation with the workgroup for the types of concerns that you're raising.

Member Chin: I guess my main concern is not so much the penalty to the clinician, but the consequence to the patient. So if you risk-adjusted, and we're not -- the physicians aren't seeing a penalty -- similarly, you know, I think there's other pressures which shift these surgeries outside of the inpatient. You know, how are we understanding what the shift is doing to quality?

Chair Martin Anderson: Yeah, can I just add on, for one thing? I struggle -- I'm okay on the use. I struggle a little bit with usability, given there's no information provided that talks about how this

measure could be improved. And always, with cost measures, what you want to protect against is that the best way to approve is to not give needed services. Right? You could get lower costs.

Now, I don't think that most clinicians would be motivated to do such a thing, because they're obviously not trying to optimize their cost measure. That's not an issue in our U.S. health care system.

But how do we think about the usability criteria, when the answer is, there's no data to demonstrate improvement? I don't know, Matt, if you have an answer to that, but -- we just rate it as low and move on? I mean, I'm struggling with how the staff got to moderate, when the answer was, we have nothing.

Mr. Nagavarapu: So this is Sri from Acumen. Yeah. I had a couple thoughts on that, and feedback you all have would be great to take back to CMS on it. And I think it's related to the question from Dr. Chin, as well, about quality, in terms of the inpatient versus outpatient.

It's sort of -- the measure is currently constructed, if a patient were to suffer greater complications or have a higher likelihood of complications by moving their case to outpatient inappropriately, our measure would reflect that. But that, again, ties in to your plan, Dr. Anderson, of trying to understand what's driving the measure, and clinicians being able to act on it.

And just two quick thoughts on that. In our field testing reports that are sent out to all clinicians, what we did was break down the cost performance into distinct categories of performance, like by particular complications, post-acute care use, things like that.

And then for each category, the costs of the clinician, the surgeon, was compared to both a national average and a set of providers with a similar risk composition as you, to give you a sense, not just whether you're high or low on cost, but are you high or low on post-acute care? Are you high or low on

imaging? Are you high or low on complications? To try and tease this out a little bit, and make the information more actionable.

Like, we're not ultimately on the production side of the QPP, where -- the contractor that creates all the reports that are shown to clinicians for the final measure scores. But we've had discussions with them in the past, and the hope is to try and filter as much of those types of categorizations into the final reports as possible.

And so if you have thoughts on what those final reports would look like, I'm sure CMS would be interested in hearing it. But there's, like, readymade breakdowns along those avenues. And we've used them in field testing.

Dr. Pickering: Kristine, if I could chime in, as well? So, great question. Yeah, the usability of measures and seeing if there's an action that can be taken by providers and the accountable entities as a result of the performance scores of the measure is definitely something we consider within our criteria.

Usability itself, for new measures, it's hard to make that distinction of, you know, has this shown improvement over time? Because the measure hasn't really been used, if it's a new measure, for the most part. So there may not be data to show that there may be improvement happening on the measure.

And in addition, you know, some of the unintended consequences of its use, or its potential harms, all of that information may not be readily available for new measures. Again, speaking for new measures, in which case -- these measures are. So if there is some variation that we see, and there potentially could be opportunity for actionability on that variation, this is where NQF comes to the decision of a moderate rating.

And it is still up to the standing committee to determine whether or not there is some actionability, and maybe even just a plan that there could be a use

for these performance results, in which providers or accountable entities can act on them. Then you can rate usability accordingly.

But that's where the moderate rating comes from, from NQF, is -- it is a new measure. We understand that it's not really been fully used, to the extent to where we're seeing improvement results, but there is potential for some actionability on those results. And this is where we get the moderate rating for this measure and other measures.

Chair Martin Anderson: Thank you. I think where I struggle is -- we actually create measures for broad public reporting use and payments, that we're testing in the field, as opposed to what used to happen, or could happen, which is, you create a new measure and you actually test it in the real world, before you put it in a public reporting program or incentive program. But I think that's where we've evolved to, and it's certainly not the fault of the developers that that's where we are --

Chair Jhamnani: -- which is what I was going to ask the developer, in terms of -- because based on how we've dealt with NQFs in the past, most of these reports are presented when the measure's up for maintenance endorsement.

And this is where I wanted to ask the developer if they have any ideas as to -- how did it improve? Because there's certain challenges. I mean, if you look at field testing reports generated, the response rate is very low with each report. A lot of clinicians don't even understand the math that goes into this. They don't understand what these numbers mean, how to make clinical sense of these numbers.

And then there are not a lot of useful, actionable data within it. The only thing is saying, yes, these certain procedures, for me, cost more, but then I'm trying to figure out, where is my area of improvement? Is that because I've put in an expensive prosthesis? Or are there more complications? Or are there more -- they're going to the wrong place?

And those information may be there or may not be there, or may not be understandable. So I think those are things that Acumen and the CMS needs to keep in mind, if the ultimate goal is to make this actionable, when it comes for maintenance.

Member Gidwani: Yeah, agreed. I agree with that sentiment, and also wanted to bring up my past question. I think now is the time that we can talk about it. And it sort of piggybacks on this discussion, which is, do physicians have the ability to move the needle on this? Especially those, you know, half of physicians in the country who are employed by hospitals and health care systems.

So in that vein, I think it would be very helpful for us as a review committee to be able to understand, what are the cost contributors? You know, where are the majority of the differences between the ends of the distribution coming from? Is it coming from imaging? Is it coming from post-discharge care? Is it coming from readmissions? Sort of helping us understand that, I think, will be elucidating in letting us know whether the measure's aimed, again, at the right actor.

Member Chin: I mean, I think if we look at CJR as a, like, predecessor, where you gave physicians or hospitals the opportunity to try to create cost savings across the episode of care for joint replacement, most of the savings came from just not using institutional post-acute care. Like, there's not that -- like, I think the cost savings for everything else was so much smaller. So this measure becomes really focused on, like, that one piece, is my impression.

Where this measure is different, though, is, it's not just impatient. Right? It's also looking at all these other settings. And I know that we said that we adjust for it, but I think the behavior and structures within each of those settings is slightly different. Right? So we might want to think about what that means.

Chair Jhamnani: And your point is valid, Amy,

because if you look at most clinical processes, there are just few low-hanging fruits, where you can cut costs. But after that, where does the commission cut costs? And it's not clear.

Like, so -- and that's the struggle that I have, as a clinician, with many of these measures, because, as you said, we can decide that the patient doesn't go to post-acute care, and just goes home, or something like that. But after that, then what?

Chair Martin Anderson: Thank you. Pam?

You might be on mute, Pam. You'll get it by the end of the day.

Member Roberts: Yeah, sorry. It's still early here. I think that this -- it will push the clinician to be able to partner with institutions, with post-acute care, and other entities within the community, that it may have some benefit. I know from BCPI, having participated in that, it really pushes a community to work together. So there is some benefit in that.

Chair Martin Anderson: Thank you for that perspective. Okay. We do not have to cover Criterion 5, so I'm going to give one last pass to see if there is anything else on Criterion 4, on use and usability, and you can make your notes.

I also want to make everyone aware that Joyce did put some notes in the -- further comments in the chat, related to the mortality question, which I think you guys can read before you finalize your validity. I don't think we want to reopen the whole discussion again. I think it's clear, what she put in there. And Joyce, I see you have your hand up again. What did you want to discuss?

Ms. Lam: Oh, sorry, I was just piggybacking off that last conversation, about cost savings. But I can go after you talk about the validity.

Chair Martin Anderson: Okay. Is there something specific you wanted to -- I mean, is it a direct answer

to one of the questions, Joyce?

Ms. Lam: Oh, it was the questions about what clinicians can do to move the needle. So as Dr. Chin says, the CJR showed cost savings. A lot of it came from PAC, but our measure savings can come from other things besides just PAC. And so, looking at -- the correlation of complications was a .28. It was very promising.

And in terms of other services, that's something that -- you know, the process we're developing for identifying all the types of services that are clinically related, so that clinicians, when they receive this information, can review the intensity of the frequency in practice patterns, to make adjustments to other parts of care, not just post-acute care.

Chair Martin Anderson: Joyce, can you clarify what you mean by, that was encouraging? Because a 28 percent correlation to me would say 70 percent of the time you have a complication that's not correlated to higher cost. Is it a subset of complications that you think is promising for making changes? Or is it just the fact that there is some correlation?

Mr. Nagavarapu: Hi, Dr. Anderson. This is Sri. Yeah, I think what we're using as a benchmark is just the correlation with quality measures that we'll typically see for NQF submissions, which can kind of be all over the ballpark, from .05 to .5. And so the .28, or close to .3, is high, relative to lots of the correlations that we typically see.

The reason that I think of that as encouraging is because of the comment that was made earlier by Dr. Chin about the fact that a lot of the opportunity for improvement in lower joint replacement, historically, has been on the post-acute care side. And we know, even in our measure, that some of the opportunity for improvement is on the post-acute care side.

But given that there are really big-ticket items here, like post-acute care, imaging -- that definitely differs across clinicians -- and potential follow-up services

associated with rehab and so on, the fact that complications as a category is picking out a correlation of .3 seems encouraging from the point of view that there's a lot going on in the measure. There's a lot of cost drivers here.

And my worry, sort of approaching the measure agnostically, initially, would've been that, you know, all of the cost-driving and the savings would've been coming from post-acute care, and like, that at .28 is a sign that it's not and, I think, is a good sign. And, you know, to the extent that some savings can come from complications, some savings can come from reducing excessive imaging, and so on, is important, and a good finding.

The other aspect of this that was encouraging is just, like, that correlation sort of, in some sense, masked very large cost differences for having the complication versus not. So the incentive to not have a complication is extremely high.

So if you look at the mean observed of cost with hip complications, it's around $29,000, and the mean observed costs without hip complications is around $18,000. And so embedded within the measure, there is a -- it's a pretty large incentive to avoid complications, as well.

Chair Martin Anderson: Thank you. Okay. Any closing comments from anyone before we put our votes in on SurveyMonkey, break for lunch, reconvene at 12:30?

That's a tall order, right, to throw a comment in that takes away your extra ten minutes for lunch? Okay. So we reconvene at 12:30, and we'll continue with Measure Number 3626. Or, no, 3625. Thanks, everyone.

Ms. White: Wonderful. Thank you. See you all at 12:30. Thank you.

(Whereupon, the above-entitled matter went off the record at 11:49 a.m. and resumed at 12:31 p.m.)

Ms. White: Okay, we will go ahead and we'll get started. And shortly, we will be sending an email to the entire standing committee, the participants on the call today, that will include the three attachments, the three tables, that are -- for each measure, 3623, 3625, and 3626, that were provided by the developers. So I will let everyone know when those emails are sent out, so that you can look out for those.

And then we will just pause here for a moment to pull up our slides, so that we can showcase our next measure under review. So thank you so much. Wonderful.

So our next measure will be Measure 3625, Non-emergent Coronary Artery Bypass Graft Measure, and Sunny will be leading our discussion for this measure, so I will turn it over to Sunny.

Chair Jhamnani: Thank you, LeeAnn. Good afternoon or good morning, wherever you are. This next measure and the next measure after that are pretty much -- have a lot of similarities to the first measure that we talked about, so many of the concerns or questions should have already been addressed, so these two measures should go fast.

This measure is 3625, Non-emergent Coronary Artery Bypass. We cardiologists and physicians like to short-form a lot of things, so we call this as CABG -- for persons who aren't familiar with the verbiage -- Measure. The measure steward is CNS and Acumen. And LeAnn, can you share back the screen? I can't see that.

Ms. White: Yes. Apologies. Victoria, can you share the screen again? Thank you.

Chair Jhamnani: Thank you. And this is pretty much an episode-based cost measure which evaluates a clinician's risk-adjusted cost to Medicare for patients

who undergo CABG procedure during a performance period. Measure score is the clinician's risk-adjusted cost for the episode group, averaged across all episodes attributed to the clinician to the commission.

This procedural measure includes cost of services that are clinically relates to the attributed clinician's role in managing care during each episode from 30 days prior to the clinical event that opens or triggers the episode, through 90 days after the trigger.

Patient populations eligible for the Non-Emergent CABG Measure include Medicare beneficiaries enrolled in Medicare's Part A and B. I'll open up to the developer, to see if they have any other comments or would like to give any additional overview on this measure.

Ms. Lam: Thanks. So as you mentioned, this measure shares a lot of features with the hip arthroplasty measure, so I'll just focus on the specifics of this measure.

So the focus of this is the inpatient care for CABG. And we've focused specifically on non-emergent CABG procedures, so the trigger logic and exclusions do reflect this. So for example, it excludes episodes where the principal diagnosis on the inpatient claim is for STEMI. The episode window is the same as the hip arthroplasty, so 30 days prior to the trigger, and it ends after 90 days.

Some examples of the services that are included in this measure are coronary disease readmissions, wound care, imaging, preoperative workup, and other types of cardiovascular care.

The attribution and the way that the score is calculated is the same. For specific risk adjustments that speak to the clinical risk factors -- so this was something that we worked with clinical experts on. So this includes specific conditions and other risk factors like anticoagulant use, recent hospital admissions, antiplatelet use.

The development process -- the measures went through the same structured process where we did iterative development and testing. There were over 60 clinical experts involved in the development across two panels. And we sought input from five individuals who shared their lived experience of CABG.

For field testing, we produced over 3,500 field test reports for this measure. And like the hip arthroplasty measure, CABG was added to the MIPS cost performance category in 2020. And while the CABG measure is not yet included in BP, there are several CABG outcomes measures in the MIPS quality performance category, which offers some really strong opportunities for alignment.

These include measures for prolonged intubation, and there's also a risk-adjusted operative mortality measure for CABG. So that's something that speaks to the concern about the -- how we exclude episodes ending in death, where mortality will be assessed through this quality measure that's designed to look at the risk-adjusted mortality rate. So these costs and quality measures will work well together in MIPS to holistically evaluate the value of care overall with CABG.

I also wanted to just highlight a couple of opportunities for improvement, since that was something that came up with hip arthroplasty. So clinicians who have higher costs can explore a number of areas for improvement, including reducing readmissions, reducing complications, and also assessing and planning for higher risk factors preoperatively, and appropriately ordering tests before and after the surgery.

Some examples of interventions could include preoperative education and early discharge planning for high-risk patients, in particular, to avoid readmissions, and following guidelines for recommended imaging and testing, both before and after the surgery.

We also heard from our Person, Family and Caregivers partners, who identified areas where the care from the attributed clinicians was particularly impactful, including the importance of discharge planning in facilitating their recovery. Some examples that these partners shared with us was learning about wound care, learning about taking medications, and also, how to identify complications, such as recognizing blood clots.

Some also noticed the impact of coordination between the surgical team and specialists treating co-occurring conditions as part of discharge planning. So these are all just some examples of opportunities for clinicians to improve their performing on this measure.

Chair Jhamnani: Thank you, Joyce. Our lead discussants -- discussant, I would say, because Dr. Kalra, I'm not sure is present. Right, LeAnn? Dr. Kalra is not present today?

Ms. White: I do not see him on our participant list, Sunny.

Chair Jhamnani: So it will be Danny. Danny, my man.

Member Van Leeuwen: Okay. Yes, well, I'll do the best I can here.

Chair Jhamnani: Don't worry. I'll help you out.

Member Van Leeuwen: Thank you. I kind of got stuck on Number 1, because -- the impact of high-resource use. So I was immediately interested that the study referred to, a period from 2000 to 2012, that an average of 100,000 Medicare patients underwent CABG, each surgery, each year. And 12 years is a long time for an average.

And so I looked at this source document, and what it showed is that there's been a steady decrease in volume of CABG over those years, as well as a steady decrease in mortality.

So then I went to see, was there any more recent data? Because 2012 is ten years ago. And there was a study -- and I can't remember what it was -- I think it was referenced in this document -- that was from 2003 to 2016, which also pointed out the steady decrease in incidence of CABG and mortality.

So it makes me wonder, are we just on a downward trajectory of incidence and mortality, just as a matter of course? And what would be our ability to improve if it's all decreasing anyway?

Chair Jhamnani: I think, Danny, there are a couple of questions embedded in that. Overall trends and procedural volumes for CABG have been decreasing. And this has panned out over experience and data, as you've pointed out. It's due to changing landscape of patient morbidity, advancements in interventional cardiology, which have allowed many of these procedures to be done percutaneously.

Improvements in mortality is a good thing. But your point is, have we reached a plateau, where there's no point of improvement or very little scope of improvement? And I'm not sure -- I do think that there's scope of improvement. That is my personal take on that, as a provider. So those are my two cents, if any of the committee members have anything to add to that. Does that answer your question, Danny?

Member Van Leeuwen: Does it answer my question? I think -- I guess I would be interested in hearing what the developer, what Acumen says. You know?

Chair Jhamnani: Sure.

Member Van Leeuwen: Yeah.

Chair Jhamnani: Sure. Joyce?

Ms. Lam: So in terms of number of episodes, about 41,000 for 2019. The mean observed cost is $43,000. So there's a lot of -- it's a very high-cost area. And this is in the testing attachment, which I think you

should have just received. So it's on the tab noted Table 2b2.2, Exclusions. So I'm looking at the final rows on that table.

So it's -- and that's one of the reasons why we developed this measure, is just -- it's a lot of costs for Medicare, and -- with the areas for cost improvement that I mentioned and which Heather, a clinician on our team, can speak more to in detail if you're interested in hearing some more examples.

And in terms of mortality -- so our measure really just focuses on cost. So we haven't been tracking changes in mortality over time. But just noting that the way MIPS works is that the cost and the quality measures work together, and there is a risk-adjusted mortality measure specifically for CABG.

Member Van Leeuwen: Thank you. I think I need to also say that the staff rated this as moderate. Anyway, I have another question, but I don't know whether it belongs here or in Criteria 2. So let me say it, and then you can tell me where it belongs. One of the things that I see interesting, also, in some of this source material is that women only are -- 33 percent of the CABGs are women, and the, you know, the two thirds are men.

And so it made me ask the question, do blockages -- really, do only 33 percent of blockages occur in women, which -- I think that's probably not true. Like, so is there something -- is there something to think about in this measure that -- is the important cost consideration that CABGs are definitive treatment for everybody? Or they're not necessarily so?

And is it as necessary for men, since more men have the procedure than women? I don't understand where the, you know, where that disparity -- now that I'm talking about this, I realize it probably belongs in Criteria 2. But guide me.

Ms. White: You are correct, Danny. That's more for validity, so we can table that for the validity

discussion.

Member Van Leeuwen: Okay. All right.

Chair Jhamnani: Yeah, yes, Danny, that falls under validity. We'll get to that.

Member Van Leeuwen: Okay.

Chair Jhamnani: And any other things that you found out during your analysis, Danny, before opening to the committee members? No? Thank you. Risha, I see your hand raised, though.

Member Gidwani: Thanks, Sunny. My question is, when I was reading the background, you know, it seems to be that the opportunities for improvement were in avoiding readmissions, which are very costly, both in unit costs as well as overall costs, because so many people are getting this procedure, and also in use of appropriate post-acute care.

And so that sort of begs the question to me -- is just sort of, what was the rationale for making this a cost measure? You know, why not make this a quality measure and have this focus on 30-day readmissions or provision of appropriate post-discharge care?

Ms. Lam: So the short answer is that the episode-based cost measures are required by statute. So under MACRA, which created MIPS, it's stated that the cost performance category needs to have measures that are based on care episode groups. So really, we're fulfilling the statutory mandate to create cost measures.

And the starting point was looking at high-frequency, high-cost areas of care, which is how we prioritized CABG for development. There are related quality measures, so that was another factor that we considered, in terms of thinking about what measures to develop.

So there are four CABG-specific quality measures in MIPS 2022. There used to be more, but they've

gradually been removed. But there are still four high-priority outcome measures for prolonged intubation, postoperative renal failure, surgical re-exploration, and the risk-adjusted operative mortality measures that I'd mentioned before.

So it's really a nice illustration of an area where the cost measure works alongside these four really robust outcomes measures, which isn't available for every condition, but it can holistically evaluate all aspects of -- well, big aspects of care for the value of CABG.

The other point, to address the readmissions, is there is a hospital-wide unplanned readmissions measure, which is based on claim status. So that one is also calculated automatically for all MIPS participants. And maybe when we get to validity, we can share the correlation analysis that we did, similar to for the hip arthroplasty measure.

Chair Jhamnani: Thank you for that, Risha and Joyce. Any other questions on importance to measure and report?

Okay. So we're going to go next to the scientific acceptability. We're going to start with reliability. Danny?

Member Van Leeuwen: Well, thank you. The one thing that I had remembered from our previous session six months ago was this, you know, looking at the mean reliability being above .7, and I see that it is, for this measure.

So not that I really know what that means, but I see that it's there, and that the staff rated it as moderate. Otherwise, I really don't have anything else to say about reliability. Somebody sure could help me out, if there is more to be said.

Chair Jhamnani: Thank you. Thank you, Danny. I'll open up to the committee members. You're right about the threshold that we discussed during our last meetings. I'm glad you caught that. It was a point of

contention during our last calls. Any questions about reliability rating, reliability methods, or anything like that? And LeeAnn, go ahead. Is there some --

Ms. White: Yeah. Sunny, if it's okay, I would also like to just recap what we have in our measure evaluations worksheet document for reliability and the specifications. So I just want to quickly review, again, that this measure was reviewed by the SMP, and they found the measure to be reliable. They did pass the measure with a moderate rating for reliability.

The developer did conduct two tests for reliability performance. So they used the signal-to-noise analysis method and the split sample analysis. For both methods of calculation, the reliability was calculated for TIN and TIN-NPI, and had a case minimum of ten episodes.

So as Amy said, a signal-to-noise test found that the mean reliability for TINs a 0.84, and for the TIN-NPI, the combination is 0.75. The one examined by the number of clinicians was in a practice, the average reliability score did increase from 0.76 -- that was one clinician -- to 0.97, 21-plus clinicians for the TINs level.

And then for the split-sample testing, the ICC or intra-class correlation coefficient was 0.80 at the TIN level, and 0.64 at the TIN-NPI level. So I just wanted to go ahead and quickly recap on those, as well. Okay.

Chair Jhamnani: Thank you. Thank you, LeeAnn. Very robust at the TIN level. Maybe not -- slightly less at the TIN-NPI level, but no major red flags for me here. Risha, go ahead, please.

Member Gidwani: Thanks. One more question I had, and this is sort of global across all three measures, is the threshold of ten cases or more -- I'm wondering why that threshold was determined. Why not maybe something higher, like a threshold of 30?

I noticed that you're using OLS in all of these models, and so, you know, relying on central limit theorem with an N of 30 or greater seems important. Can you walk me through a bit of the rationale for this lower threshold of ten?

Ms. Lam: Sure. So the process of determining the case min is something that we go through with CMS. And one of the really important considerations is weighing up coverage with reliability, because increasing the case minimum will reduce the number of clinicians who can be assessed under this measure.

And thinking about how the MIPS cost performance category works as a whole, there's a number of these episode-based measures in 2022, there's 23, and there are two population-based measures. One is the Medicare spending per beneficiary clinician measure. So that applies to all the patient care.

And one of the things that we want to make sure is that when we develop these clinically specific episode-based measures, that smaller providers, who will provide us with lower volumes of cases, are able to be assessed on these clinically-specific ones, and not just on the global population-based measures.

So the process -- we look at the tradeoffs between coverage and reliability. And so based on the testing for this measure and for the other measures, looking at the distribution of reliability, it's very high. So with a case minimum of ten, since the reliability was sufficiently high, and is far above the standard that CMS established through rulemaking, we just went forth.

That's how we came up with the case min of ten episodes. And when the testing shows that the case minimum needs to be higher, then it can be higher. So for instance, there are a number of measures where the case minimum is 20 or 35. So it really depends on the testing.

Member Gidwani: Can you tell me -- I didn't see it

here. Maybe I missed it, because I know you included it in another measure -- whether there was a difference in ICC or the signal-to-noise ratio when you were looking at providers who had, let's say, ten to 30 episodes, versus 30 or more episodes?

Ms. Lam: We looked at a distribution of reliability scores by practice size. So we looked at the number of TIN-NPIs within a TIN. So we did that for signal-to-noise. And so what we found is that the -- this is in Table 2 of the testing form -- that the mean is .84. The mean for tens was 2 to 4, TIN-NPIs is .81. For TINs with five to 20 TIN-NPIs, it's .88, and then with over 21 TIN-NPIs it's .97.

Member Gidwani: Thank you.

Chair Jhamnani: Thank you, guys. Any other questions on reliability?

Okay. Thank you. We will move next to validity. LeeAnn, can you summarize the findings of the committee analysis? And then I'll hand it over to Danny.

Ms. White: Absolutely. Thank you, Sunny. So the attribution for this measure -- the measure is attributed to clinicians and clinician groups. The approach was developed to fairly evaluate clinician cost performance for non-emergent CABG procedures and promote efficient and high-quality care for Medicare patients undergoing these procedures.

The cost approach -- the developer noted that the non-emergent CABG episode-based cost measure evaluates a clinician risk-adjusted cost to Medicare for patients who undergo a CABG procedure during the performance period. This measure score is the clinician's risk-adjusted cost for the episode group, averaged across all episodes attributed to this clinician.

This procedural measure includes cost of services that are clinically related to the attributed clinician's

role in managing care during each episode from 30 days prior to the clinical event that opens or triggers the episode, through 90 days after the trigger.

The measure also uses Medicare standardized pricing to payment standardized the Medicare claims that -- the developer did provide a link where you can download that methodology, as well, in the submission forms.

For validity testing, that was conducted at the performance measure score level. Developer conducted both empirical validity testing and a systematic assessment of face validity through a measure-specific expert panel called the Non-emergent CABG Clinician Expert Workgroup.

Out of the nine workgroup respondents to the survey, all nine 100 percent agreed that each of the measures that -- or specifications helped the measure capture clinician cost performance as intended. And eight -- so 89 percent -- agreed that the scores from the measure as currently specified provide an accurate reflection of clinician cost effectiveness.

The developer furthermore also evaluated the empirical validity of this measure by examining correlation with an NQF-endorsed measure of resource use. So that measure is NQF 2158, Medicare Spending per Beneficiary Hospital, which assesses the risk-adjusted cost to Medicare for services performed by hospitals and other healthcare providers during an MSPB hospital episode.

The developer posited that the Non-Emergent CABG Measure cost score for a provider would be consistent with the performance on that measure, the NQF 2158. To assess the consistency, the developer analyzed distribution of Non-Emergent CABG Measure consistency -- or scores. I'm sorry. Apologize. And then -- which is, they observed the expected cost ratio across MSPB performance ratings.

They saw that the MSPB performance ratings increased or improved. The mean observed-expected ratio of a Non-Emergent CABG Measure decreased from 1.03 to 0.98. so the cost and performance improved, as hypothesized.

I am going -- they also did clinical inclusions and exclusions. Do you want me to pause before we get into the exclusions, Sunny? Or do you want me to --

Chair Jhamnani: Just summarize everything, LeeAnn.

Ms. White: Okay. Perfect.

Chair Jhamnani: Thank you.

Ms. White: So the developer examined the distributions of observed costs and ratio of observed over expected spending calculated by applying existing risk factor coefficients to exclude the episodes for each excluded population. They then compared the cost characteristics of the excluded episodes to that of episodes included in the measure, to assess the distinctness between the two patient cohorts.

The developer provided data on the observed cost statistics, and observed-to-expected cost ratios for the Non-Emergent CABG Measure exclusions. The cost statistics are also provided for the episodes included in the measure for comparison with a ten-episode case minimum at the TIN and TIN-NPI level. The statistical results provide evidence that excluded episodes are not comparable to the overall measure population.

For risk adjustment, the developer controlled for case mix using a statistical risk model with 110 risk factors. The risk adjustment model for the Non-Emergent CABG Measure broadly followed the CMS Hierarchal Condition Category risk adjustment methodology that's used in the Medicare Advantage program. This model includes 79 indicators, as specified in the model, derived from the patient's Parts A and B claims during the period 120 days prior

to the episode trigger.

The developer also used the CMS enrollment database and common Medicare environment to determine dual eligibility rates effects. Socioeconomic status was determined by two approaches. So they used income, education, and employment status as categorical dependents, and they used the Agency for Healthcare Research and Quality, AHRQ, SES Index as a continuous dependent. Both approaches used the 2017 American Communities Survey by linking episodes to census block groups, ZIP code and -- when the census block group is missing.

Social risk factors were also examined, relative to the base set of risk adjustment variables from the CMS HCC 2015 model disability status, end-stage renal disease status, interaction variables, and recent long-term care use, and in a stepwise fashion to determine the social value of each social risk factor that's considered. The developers also analyzed race, sex, dual status, income, education, and unemployment as social risk factors.

They examined the impact of including social risk factors into the risk adjustment model by running goodness-of-fit tests when different risk factors or added and compared to the base risk adjustment model. The developers did note that the results from the stepwise analysis do not support the inclusion of social risk factors into the Non-Emergent CABG Measure risk adjustment model.

The overall R-squared to the Non-Emergent CABG Measure, calculated by dividing explained sum of squares by total sum of squares, is 0.442. The adjusted R-squared is 0.440.

For the meaningful differences, the developer used two methods to identify statistically-significant and meaningful differences in the Non-Emergent CABG Measure. First, they analyzed the distribution of performance scores for the overall measure, as well as for clinicians, stratified by meaningful provider

characteristics. So urban versus rural, census division, census region, and the number of episodes attributed to the clinician.

They also calculated the 95 percent confidence intervals, using the variance of the provider mean, and then the compared each clinician's 95 percent confidence interval to the average national measure score, in order to determine if the clinician's performance was significantly different form the national mean.

So the developer did note that the Non-Emergent CABG Measure scores have a good deal of variability. For the TIN level, the standard deviation is .09, and the 99/1, 90/10, and 75/25 percentile ratios are 1.48, 1.20, and 1.09, respectively. And then at the TIN-NPI level, the standard deviation is .08, and the 99/1, the 90/10, and the 75/25 percentile rations are 1.48, 1.21, and 1.10, respectively.

We do have exclusions. In general, I do need to mention the SMP Subgroup members found the measure to be valid, and they passed this measure with a moderate rating for validity, but they did have some concerns, regarding the empirical testing, that I'll list for you today.

First, their first concern is, there's not enough variability between the MSPB performance rating category to show meaningful information. They also would like to see clarification of how costs are narrowly defined to only those associated with CABG, and its post-acute care.

The developer did provide the following responses to the SMP concerns. They list the clinically-related services that are detailed in the Measure Code list file that we did provide you. And I just want to let you know, we -- please check your email. We did send the table for your reference. The Clinical Expert Workgroup members reviewed analyses of the utilization and timing of all Medicare Part A and B services, relative to episode trigger, to identify services for inclusion.

Five patient representatives provided input through structured interviews, based on their experiences of undergoing a CABG, and this perspective is reflected by including the care services that patients experienced, including imaging, testing, wound care, cardiac rehab through different types of PAC, and follow-up visits with the surgeon.

And lastly, for the exclusions, it is a major concern that almost half of the episodes are excluded at both levels, TIN and TIN-NPI levels. Another concern pertains to excluding patients who may potentially die during the episode duration, due to low quality of treatment.

Several SMP members raised concern about the risk adjustment model. There were questions about not including social risk factors, especially the dual eligibility status, in the final risk adjustment model, and the clustering of factors associated with homebound status, and when they constitute unmanageable clinical risk. So that -- I will hand it over to Sunny and Danny to lead us in the discussion of validity, and add on to that. Thank you.

Chair Jhamnani: Thank you, LeeAnn, for that comprehensive overview. Danny? I can't hear --

Member Van Leeuwen: I'm on mute. Yeah. Thank you. I want to go back to -- is being female a risk factor here? Because if it's true that only a third of CABGs are performed on women -- and I would assume that there's probably not just a third of, you know, a third had blockages, you know, I mean -- so there's something different about how we are treating men and women with CABGs.

And you know, I don't know how that fits into this structure, like, if it's irrelevant because -- for whatever reason, or it's relevant, because there's some real disparity here. And this is either an underutilization for women, or an overutilization for men, or -- and how does that affect cost? And how does that affect risk factors? So I need some help in giving this some context in our evaluation structure.

Chair Jhamnani: Sure. Joyce, have you done some analyses that can answer Danny's question?

Ms. Litvinoff: Hi, this is Heather. I'm one of the clinicians on the Acumen team. I'm just going to jump in, just to give some additional context. So the numbers that you see in this measure are similar in the literature. And I just do want to point out that there's a number of gender-based anatomical and physiological factors that are different between men and women that may be contributing to this.

Some of these include things like smaller coronary artery size. There is heart failure with preserved ejection fraction, more commonly in women. There's the role of post-menopausal estrogen withdrawal, which has a role in arterial sclerosis.

Also, it's been cited that there may also be some differences in these numbers due to delayed onset, and just the ambiguity of symptoms of coronary artery disease, which can lead to delayed diagnosis in women. So, just wanted to provide a little additional context with respect to the gender differences. Hopefully that's helpful.

Chair Jhamnani: Thank you, Heather. Joyce, do you have any numbers in your analysis? Or is that the only thing that you could toss us?

Ms. Lam: We're just looking through this for anything that could be helpful for you guys.

Chair Jhamnani: Okay. Danny, it's a very valid point. Gender disparities are something that we, as a cardiology community, look at. And as I think Heather mentioned, there are many reasons for that. And I'm not sure there's one unifying answer that can answer that. It's complicated, in a nutshell. But thank you for pointing that out. It's definitely a concern that we should be thinking of. Risha, do I see your hand raised?

Member Gidwani: Yeah, I'm looking at these Excel tables, and I'm confessing, I'm having a hard time

understanding what is actually in the final model. So I see on Table 2b3.4b, the social risk factors -- but there's a lot of models that were tested. Which is the final one that's being used here? Is it Model 1? Model 2?

Ms. Lam: Yes, it's Model 1, the base model.

Member Gidwani: Model 1. Okay, and so that excludes sex. Is that correct?

Ms. Lam: Yes, that's correct.

Member Gidwani: Okay. So I think, then, Danny, that this actually gets at your concern. We're essentially, then, telling providers that it doesn't matter whether you're treating a male or a female. We still expect you to have the same type of resource use across both sexes.

Chair Jhamnani: Any other concerns or questions on this issue?

Member Van Leeuwen: Not from me.

Chair Jhamnani: I do have a question. I mean, some of the questions pertained to my questions from the prior discussion, and I won't bother you again. I would cite what the SMP mentioned about the correlation with the MSPB measure.

One of the things that I wanted to bring up here is the exclusions. If half of the events or procedures were excluded, what does that do to the disparities within the core data that's already included? Do you have any sense of that? Does that change the dynamics in any other way on certain aspects that we are missing along the social determinants of care, even though we did not include them?

Ms. Lam: So this is actually in the testing appendix, on the first table. So 1.6, Inclusions. So we looked at different demographic characteristics, and the inclusion and exclusion criteria, to see if we're distorting anything in our patient population through

those exclusions. And what we see across the characteristics is, it's pretty similar from inclusion to exclusion criteria.

And for the point you mentioned about how we exclude about half of the initially-treated episodes -- so part of that reflects data cleaning. So we require continuous enrollment in Medicare Part A, B, and not C, so that we have complete data for all the episodes. So we do drop a lot of episodes for data completeness reasons.

And the way this triggering logic was designed was to focus just on non-emergent CABG procedures. So we do have exclusions that are designed to make sure that we're not accidentally picking up emergent CABG. So that's another one of the factors going into how we came up with the exclusion criteria, and why the number of episodes is different.

And the last one, of course, is case minimum, so episodes where the provider doesn't meet the case minimum, they ultimately aren't in that final row that you might be looking at.

Chair Jhamnani: And -- go ahead, please.

Ms. Lam: Oh, the other point that you mentioned, about those correlations -- so as I mentioned, we did construct the unplanned readmission rate so that there's another piece of information that you can consider for thinking about correlation to quality measures.

And so what we found is that, similar to the hip arthroplasty, we saw that there's a medium to high correlation. So Pearson correlation at the TIN level is .35, and at the TIN-NPI level is .1. So it reflects the relationship that we expect, which is that clinicians who do better on the cost measures tend to have lower rates of unplanned readmissions.

And the unplanned readmission rate -- sorry, I'm using the short name. It's the MIPS quality measure, the hospital-wide 30-day all-cause unplanned

readmission rate for MIPS. So that's a measure that is a really specified version of the all-cause readmission measure. So it's based on a well-established measure that's in use.

Chair Jhamnani: Thank you for that, Joyce. The one question that I had is, you have AVR there, but you don't have other valves. Was there a reason for that? Or what why did you just include AVR, aortic valve replacements?

DR. DO: Hey, this is Rose. I'm one of the cardiologists at Acumen. I apologize, I was at another commitment, so I'm joining a little bit late. I was familiar with the workgroup that chose the specifications, and I think -- we based a lot of the decision-making on numbers.

So I think this also ties back to the gender disparity that we see within cardiology itself. I think there's referral bias, there's a lot of complicated reasons, I think the study mentioned, about why we're seeing more.

The other thing that led us to decide on what valve to choose was also based on numbers. And I think it's just more that, you know, you see a lot of aortic stenosis and CAD together. And so they were the big buckets that we decided were clinically distinct enough that we wanted to do some subgrouping off of that. Of course, we did look at the other valves, but I think it all, you know, just kind of boiled down to, what is going to be the largest population that a regression is going to work on?

And I think, you know, going back to everything, the reason this episode was chosen, despite the gender differences is because it's low-hanging fruit, it's something that's a high cost to Medicare. A lot of procedures are done. I think we're going to start seeing more women being referred, hopefully, you know, because I think some of it is, again, referral bias, and then also, just the physiology that Heather pointed out.

But it means that it should be measured. I mean, we probably need to see these differences. We probably need to put a spotlight on how medicine is being practiced differently, and the cost, you know, ramifications of that. So I hope that helps.

Chair Jhamnani: Thank you, Rose. Any other questions from the committee members on validity?

Okay. Hearing none, I think we can wrap up scientific acceptability, and we can go to feasibility. LeeAnn, do you have anything to summarize for us over here?

Ms. White: Thank you, Sunny. So just to remind the standing committee, feasibility is the extent to which the specifications, including the measure logic, require data that are readily available or could be captured without undue burden to the clinicians or accountable entities.

The developer did note that all data elements are in defined fields, and that a combination of electronic sources and -- the measure uses administrative claims data. The preliminary analysis or rating by the NQF staff for feasibility was rated as high. I'll hand it over to Danny, if he has anything else to add to that.

Member Van Leeuwen: Nothing to add. Thank you.

Chair Jhamnani: Thank you, LeeAnn, Danny. Committee members, do you have any concerns on feasibility, or questions?

All right. We will move to use and usability. LeeAnn, can you summarize for us, please?

Ms. White: Yes, thanks, Sunny. So for use, this criterion evaluates the extent to which audiences use or could use the performance results for both accountability and performance improvement activities. We have two subcomponents for use that we look at. We look at the accountability and transparency aspects.

So performance results are used in at least one

accountability application within the three years after initial endorsement and are publicly reported within six years after initial endorsement. This is a new measure. If it's not in use at the time of initial endorsement, then a credible plan for implementation within the specified time frame is provided.

Feedback on the measure of those being measured or others is our second subcomponent. So we have three criteria that demonstrate feedback for the measure. Those being measured are given the performance results, or data, as well as assistance with interpreting the measure results and data.

We also look at those being measured and other users having been given the opportunity to provide feedback on the measure performance or implementation. And the third criterion for feedback would be, the feedback has been considered with changes are incorporated within the measure.

The developer indicated that this measure is publicly reported. It is currently used in an accountability program. For the accountability program details, the developer indicated that this measure is currently used in the Quality Payment Program, QPP, Merit-based Incentive System, so QPP MIPS program.

However, the developer also noted that, as specified in the calendar year 2020 Physician Fee Schedule final rule, this measure will be implemented as part of the MIPS program beginning in the 2020 MIPS performance year and 2022 MIPS payment year.

For the feedback on the measure by those being measured or others, the developer noted that during the development of this measure, the non-emergent CABG field test reports were provided to the same sample of eligible clinician groups and clinicians. Each report included information for the Non-Emergent CABG Measure if the clinician or clinician group was attributed ten or more episodes.

All stakeholders who received those reports,

including those who did not receive a field test report -- apologies, let me correct that. All stakeholders, including those who did not receive a field test report, could review a mock field test report that was listed on the CMS website.

During the field testing, the developer conducted education and outreach activities, including national webinars, office hours with specialty societies, and Help Desk support. The developers sought feedback on the reports and measure specifications through an online survey with option to attach a comment letter.

After completing field testing, the developer compiled the feedback that they received, provided through the survey and comment letters, into a measure-specific report, which was then provided to the Non-Emergent CABG Workgroup, along with the empirical analysis to inform their discussions in evaluation of any refinements that were needed to ensure that the measure is capturing what it was intended to capture.

While there were no measure-specific comments received for the measure, the group considered cross-measure feedback, reviewed updated testing results, and discussed pending items from previous webinars, voted to recommend the refinements, which were then implemented in the measure.

And lastly, for additional feedback on the measure, the Non-Emergent CABG Measure was implemented in MIPS after going through the pre-rulemaking process and notice-and-comment rulemaking.

The measure was submitted to and included in the 2018 Measures Under Consideration list. It was then considered by the National Quality Forum's Measure Applications Partnership Clinician Workgroup and Coordinating Committee in December of 2018, and January of 2019, respectively. The MAP conditionally supposed the measure, pending NQF endorsement.

MAP noted that CMS and the Cost and Efficiency Standing Committee should continue to evaluate the risk adjustment model of this measure, and consider

whether there is need to account for social risk factors in the model. So that is a summarization of the use criterion. I'll hand it over to Danny for any further additions.

Member Van Leeuwen: Thank you. And thank you for summarizing. It was beyond my ken. Anyway, the one question that I had, looking over this, is, you know, this one is, you know, how can the performance result be used to further improve health care? And I was trying to find, wherever, like, an example of a cost measure -- a value or -- the data has led to improved care outcomes, so that that's something that could be shared with the public.

And I couldn't find anything. And that doesn't mean it isn't there. It just means I couldn't find it. But I just sort of want an example of, like, well, how would we recognize that it can be used to improve stuff? Not just that there's an opportunity -- we know there's an opportunity -- but that it's been done.

Ms. Lam: So looking at the improvement over time, that's something that we do through the routine maintenance process. So as the measure hasn't actually been reported yet -- although for 2021, providers will get the beneficiary-level reports -- so we don't yet have information that we could look into, in terms of seeing practice changes or areas of improvement, since the introduction of this measure.

And as Sri mentioned for the hip arthroplasty measure, the information going out to providers in these episode-level reports provided a lot of detailed information. And over time, if we and CMS hear more feedback from providers about what would be actionable information for them to be able to identify and improve their performance, it is something that we'll keep an eye on.

And the sorts of services that are included, for the purposes of field testing, we had categorized into themes, which could be reflected in the report. Things like renal failure, preoperative workup, postoperative labs, imaging after CABG,

rehabilitation, wound care, surgical site infection, coronary disease, cardiovascular care readmissions, and pleural effusions and pericardial effusions.

Member Van Leeuwen: So does that mean that the way we're going to see the effectiveness of these cost measures in reducing costs is that we no longer need to evaluate it, because there's no more variation? Like, what are we looking for, to say that this has been successful? I mean, I get that this is new, and we don't have the data. But I don't even know what I should be looking for, if --

Member Chin: I just want to say that I would be agreeing with Danny's point, and that this has come up in past cycles, with measures that have been around for much longer, where we feel -- I mean, I think statistically, we expect some amount of variation in every measure. Right? But I think the other piece is, how do we differentiate between the natural variation of a measure and what can actually be improved on?

I think with the new measure, while I understand we don't have past data, I think that there should at least be some -- maybe even, like, driven by the tech -- areas that we identify that could actually be improved on, and the mechanisms to improve that, versus, you know, just saying, well, generally, we see there is variation.

And I think another thing that has kind of -- I haven't fully developed this idea, but, like, when we exclude the outliers, what does that do? And are we missing a huge area of opportunity when we trim off those cases? Because that might be where the real savings is. Right? Or the real waste or inefficient care lies.

Mr. Nagavarapu: This is Sri from Acumen. Yeah, thank you for both of those points. In terms of sort of what you might expect as the end-state, I guess the way that I'm thinking about the model here is what we might've seen in a program like CJR, that Dr. Chin mentioned before, where you see these cost reductions coming early on that gradually flatten out,

once you get to a point at which the certain opportunities, like post-acute care in the hip and knee example, can't go down much further.

And so gradually, over time, what I would expect is that when we internally are tracking the costs by the types of categories that Joyce mentioned, that we see a flattening out and a convergence across providers.

In terms of the size of improvement, fortunately, like, currently, where we are now, there is a large gap for improvement. You've heard some of the basic numbers on mean observed costs, and risk-adjusted cost. But just to give you, like, another look at this, if you use the Yale inpatient unplanned admission algorithm that goes into the MIPS measure that Joyce mentioned, the MIPS quality measure, and you just use that and kind of cross it with our cost measure, the mean observed cost for an episode with an unplanned admission, by that algorithm, is about $56,700. And the mean observed cost without an unplanned admission, according to that algorithm, is about $41,300. So a gap of about $15,000 off of that base of $41,000.

And so I think in the short term, there's definite places where care can improve to bring down costs, but we definitely take the point that over time, that there may be a convergence, and there may be more and more limited opportunities. And that's something we can definitely monitor over time for CMS and for you all, once we come back to maintenance.

Member Chin: Yeah, I appreciate the response, and that all makes sense to me. I think the part that I'm thinking about also is that we're not evaluating it in the context of the degree at which we do reduce readmissions. We're not even saying we've reached a level where, you know, even for unplanned readmissions, the number is so infrequent or so low that now we're not concerned about its impacting costs.

And then, what remaining cost use is there? Right? Because it's supposed to be, for this measure -- it

seems like that cost savings is so closely tied to readmissions at this point. And so in my mind, and again, this is a larger question, is -- are we setting goals? And what do we do when we reach that goal?

And again, this has come up many times where, you know, the measures become narrower, because we've been monitoring them for so long. And from that perspective, like, you know, we're scared to look away, because we don't want it to get worse, but I think even in the inception of the measures, I think we should start setting targets and goals, and then an intent of, like, the practices that we're trying to change.

Member Van Leeuwen: I like the sound of that.

Chair Jhamnani: And that is probably something that maybe CMS or even NQF can think about when we're getting these measures for initial endorsement, rather than maintenance, because these are issues that we keep on raising, and this is the biggest struggle we have, is usability.

Thank you for bringing that up, Amy and Danny. Any other questions, concerns? I'll give you an answer to -- I'll give you a moment to reply, Joyce, but let me hear from my committee members first.

Ms. White: Sunny, I'd like to just bring the comment from Risha into the recording, so that we have it on record. So Risha provided a comment related to the data that is in the Excel that we just shared with the standing committee.

She did point out that there seems to -- there seems to be a referral bias across a number of dimensions lacks -- represents 4.56 percent of patients receiving CABGs despite being -- 13 percent of the population having higher cardiovascular disease burden.

She did mention this is outside of the scope of this measure, but just applied that -- realized access to medical care could be considered in the future as an important performance measure. Risha, did you want

to add anything to your comment?

Member Gidwani: No, I mean, I think that that mainly captures it. And I don't think this is particularly in the scope of the cost and efficiency committee, per se, but, you know, maybe it's something larger for NQF to take a look at. These numbers were quite striking to me.

And so I think when we think about quality, it's important not just to think about quality of care for people contingent -- for the people who have been able to have realized access to care, but also what may be our structural barriers in the health care system, be them systemic bias, underinsurance, that may be resulting in some people not even coming in to the performance evaluation process and some patients being excluded from that, due to more upstream challenges.

Chair Jhamnani: Thank you, LeeAnn and Risha. Joyce, you had a comment?

Ms. Lam: Yes, it was just circling back to the outliers discussion. So I just wanted to clarify that the way that the measure addresses outliers is that they're excluded based on the distribution of residuals, which is the difference between expected and observed costs. So it's not just excluding high-cost episodes.

So what it actually does is, it excludes one percent of episodes at both ends of the distribution where the risk-adjustment model is unable to accurately predict costs. So outliers are spread out across the distribution of episode observed costs, since the exclusions are based on residuals. So we just wanted to offer that clarification.

Chair Jhamnani: Thank you, Joyce. Any other questions on this sub-portion of the measure?

All right. So we're done with use and usability. I don't think we have to go to competing measures. Before we wrap up this measure, I do want to open it up to the committee again, to see if there are any other

overarching issues, questions, on this measure, that they would like to bring up.

Ms. White: Sunny, would it be okay if I -- I'm just going to recap on usability. I don't think we covered usability, the unintended consequences and potential harms, unless -- someone correct me if I'm wrong.

Chair Jhamnani: Sure. Yes.

Ms. White: I know we've had a lot of discussions. So just usability --

Chair Jhamnani: It's all right. Yes. Go ahead, please.

Ms. White: Okay. So just for usability, for improvement results, I just wanted to highlight, as the measure developers mentioned, this is a newly implemented, and they did not have any updated -- or they did not provide any data to demonstrate the improvement.

And for unintended consequences, there were no unintended consequences to individuals or populations that had been identified during the testing and development of the measure. And no potential harms were identified. So NQF staff rated this as a moderate rating. And there were no concerns received from the standing committee during their pre-evaluation survey.

Chair Jhamnani: Thanks, LeeAnn. Any final questions before we wrap up this measure?

All right. Thank you, guys. I will turn it to Kristine to discuss the final measure.

3626 Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels Measure (CMS/Acumen, LLC)

Chair Martin Anderson: Thank you. So we're moving on to NQF 3626, Lumbar Spine Fusion for Degenerative Diseases, 1-3 Levels Measure, sponsor being CMS, and developer being Acumen again. It has a lot of the same, similar -- different topic area, but similar constructs.

It's a cost measure that evaluates a clinician's risk-adjusted cost to Medicare for patients who undergo surgery for lumbar spine fusion during the performance period. The measure score is the clinicians' risk-adjusted cost for the episode group, averaged across all episodes attributed to the clinician.

The procedural measure includes cost of services clinically related to the attributed clinician's role in managing care for each episode, from 30 days prior to the event, to 90 days after. Patient populations eligible for Lumbar Spine Fusion for Degenerative Diseases, 1-3 Levels measures include Medicare beneficiaries and Medicare A and B.

So, very similar in construct. Also at the clinician group practice level, clinician individual level, and multiple sites of care. So I'm going to pass it to Joyce to give the overview again. And then I believe I will be getting help from LeeAnn on this one. Is that right, LeeAnn?

Ms. White: That's correct. Yes.

Chair Martin Anderson: And our -- yes, because -- that's right. Okay. Over to you, Joyce.

Ms. Lam: Thanks. So as you mentioned, there's a lot of similarities, so I'll just focus on the aspect of this measure that I found particular to spine fusion. So -- sorry, just getting my documents in order.

So in terms of the metric construction, it's triggered by CPT/HCPCS codes for lower fusion and lumbar vertebrae in the outpatient or inpatient setting, and there's some additional conditions in the trigger logic, to make sure that we're focusing on this particular procedure.

So for instance, we exclude patients with scoliosis and kyphosis, as they often require different fusion techniques. And we also exclude episodes where the patient is undergoing a redo lumbar fusion, to make sure that we have a clinically comparable cohort. The

episode window is the same as the other two measures, where it starts 30 days before the trigger, and ends after 90 days.

The sets of clinically related services, they can generally be sort of in the following buckets. So anesthesia, pain management, wound care, thromboembolism, infection, GI complications, mechanical complication, need for revision, preoperative workup, neurological complications, and post-acute care.

The other aspects of the measure, in terms of attribution, how the score is calculated, these are the same as the other measures. For the risk-adjustment model, again, through our development process, we identified particular factors that are specific to the risk for this procedure, which includes variables such as anticoagulant use, osteoporosis and rheumatoid disease.

For the development, we went through the same process of gathering extensive clinical input, so we convened panels with over 35 clinical experts to do the measure prioritization, as well as building out all the specifications, and also engaged with four individuals who shared their lived experience of spine fusion procedures.

For field testing, we produced over 4,800 field test reports for attributed clinicians who met the volume threshold of tenets. For the alignment with quality piece, this was, again, one of the factors that we considered in deciding to develop this measure, where there are a number of quality measures in MIPS, which can assess different aspects of care.

There are four high-priority patient-reported outcome-based performance measures in MIPS 2022. So they are back pain after lumbar discectomy/laminectomy, back pain after lumbar fusion, leg pain after lumbar discectomy/laminectomy, and functional status after lumbar fusion. So again, the idea of MIPS, where you have the cost measure working with the quality

measure, the quality category, to holistically evaluate the value of care along different dimensions.

For opportunity for improvement, clinicians who have high costs can explore a number of areas for improvement, including reducing readmissions and complications, and appropriate use of postoperative PAC. Some examples are appropriate use of institutional PAC, as -- there's a study that showed that use of PAC was associated with higher odds of complications, reoperations and readmission -- and preoperative risk factor modification, which can reduce the likelihood of surgical complications, hospital readmissions, and repeat surgery. So for example, diabetes and hypertension, having these under control, minimizing the use of opioids, just some examples.

From the Person and Family perspective, we also heard, for areas of improvement or areas where clinician decisions were particular impactful in their care, some examples are engaging in counselor and care planning prior to the surgery, which also includes discussing expectations about postsurgical mobility prior to the surgery.

The individuals also shared that they felt there was room for improvement in postoperative care and rehabilitation, particularly for care coordination. And yes, that was one of the recurring themes, where we heard that they -- of an area of improvement was coordinating between the surgical care team and other clinicians after the surgery.

Chair Martin Anderson: Thank you. Sorry, I have a little trouble finding my mute button there. Okay. So let me turn it over to LeeAnn to get us started, to walk through the criteria.

Ms. White: Thank you, Kristine, and thank you, Joyce, for providing that introduction. We will start with the first criteria, importance to measure and report. Here we're looking at two sub-criteria in the high impact or high resource use, and opportunity for improvement.

The developers -- for high impact and high resource use, the developer cites that more than 6,000,000 Medicare patients were diagnosed with lumbar degenerative conditions between 2006 and 2012, and that the total admission expenditures for lumbar spine fusion surgeries exceeded $3.6 billion in 2013.

The developer posits that there are opportunities to improve the cost and quality of care related to these procedures, namely using less invasive surgical techniques to reduce postoperative complications. The measure evaluates a clinician's risk-adjusted cost to Medicare for patients who undergo surgery for lumbar spine fusion during the defined performance period.

For the opportunity for improvement, the developer provided a distribution of performance scores for clinician groups, which is identified by Tax Identification Number, TIN number, and individual clinicians -- that would be a combination of the TIN and the National Provider Identifier, or NPI -- and attributed ten or more lumbar spine fusion episodes from January 1st of 2019 to December 31st of 2019.

These scores reflect 1,415 clinician group practices and 3,330 individual practitioners, corresponding to 54,768 episodes of care for 54,768 beneficiaries. Episodes are included from all 50 states and D.C. in the following settings. Acute inpatient hospitals, outpatient facilities, ambulatory office-based care centers, and ambulatory surgical centers.

For the TIN-level score, the mean was 1.01, with an interquartile range of 0.10. for the TIN-NPI level, the mean score was 1, with an interquartile range of 0.11. The staff provided a preliminary rating of moderate. And the pre-evaluation comments from the standing committee did not reveal any concerns with this criterion. So I'll pass it over to Risha, if she has anything to add for the discussion.

Member Gidwani: Nothing from me.

Chair Martin Anderson: Okay. Any questions from the

committee, or comments from the committee, on our first criteria, on importance?

I had a quick question. Can you translate -- this is for Acumen -- the approximate spread in dollars between, say, that tenth percentile and 90th percentile on a cost-per-case?

Ms. Lam: Yes. So the mean measure score -- so on our most recent testing data of the first decile, the mean measure score is 32,000. And at the tenth decile is 44,784, at the TIN level. At the TIN-NPI level, the first decile is 31,721, and at the tenth decile, 45,093.

Chair Martin Anderson: Thank you. And the range was about the same to the lower end, too. Right?

Ms. Lam: So within the lower --

Chair Martin Anderson: Yeah, you said -- you gave the mean and the tenth. Right? The 90th is -- sorry.

Ms. Lam: Oh, yes. Sorry. This was the mean measure score within each of these deciles.

Chair Martin Anderson: Oh, okay. Got it. Got it. So the spread is about 14,000. Okay. Thank you. 12,000 to 14,000. Thank you. And is the variability also driven by -- can you tell, you know, what kind of services tend to drive the variability? Is it also radiology and post-procedure care?

Ms. Lam: So we have some of this in the same analyses that we've conducted for the other measures, where downstream acute readmissions, as well as post-acute care, have an influence on the cost, as expected.

So for example, the mean of the cost, using the unplanned readmission definition from the MIPS Yale measure that we've been talking about, for the mean observed cost with unplanned readmissions, is $53,000, and the mean observed cost for episodes without an unplanned readmission is $37,000.

Chair Martin Anderson: Thank you.

Mr. Nagavarapu: And then for other categories, you also see sort of, like, the expected correlation, the smaller -- so for instance, we have a result here for the correlation with imaging services through episodes, and that's a correlation of about .16, so small to moderate, and not as important in driving the measure as unplanned readmissions, but part of the story.

Chair Martin Anderson: Thank you. Anyone else have a question, comment?

Okay. LeeAnn, back to you for Criteria 2.

Ms. White: Okay. For Criteria 2, we will be moving into the scientific acceptability, so looking at reliability first, and we look at the specifications and the testing.

For the specifications, the measure is -- the cost measure is calculated as the sum of the ratio of observed to expected payment standardized cost to Medicare for all lumbar spine fusions for degenerative disc disease, one- through three-level episodes, attributed to a clinician or a clinician group. The resulting average episode cost ratio is then multiplied by the national average observed episode cost to generate a dollar figure.

The episode window spans from 30 days prior to the trigger day through 90 days after, and includes costs from certain clinically-related services from Medicare Parts A and B claims during the episode window. Costs are standardized to account for differences in Medicare payments for the same services across Medicare providers.

This measure was reviewed by the Scientific Methods Panel. One SMP member questioned why the measure is attributed to co-surgeons, but not other members of the surgical team. So for example, they mentioned anesthesiologists. Additionally, this SMP member also questioned why skilled nursing facility

claims were not included.

The developer provided responses to the concerns made by the SMP. Specifically, the developer noted that the attribution methodology focuses on the clinician or clinicians performing the lumbar spine fusion procedure by attributing an episode to the clinician or clinicians who bill the trigger code, which is a CPT or HCPCS procedure code. This can be both the main and assistant clinician.

They use this methodology, as the measure's intent is to assess cost related to the role of the clinician performing the surgical procedure. Since the role of the anesthesiologist or certified registered nurse anesthesiologist, a CRNA, is distinct from performing the surgery itself, this measure does not attribute episodes to members of a care team who did not bill the trigger procedure code.

Additionally, the developer noted that the measure includes a skilled nursing facility cost where the skilled nursing facility claims qualifying inpatient stay is the same as the trigger inpatient procedure. This ensures that the skilled nursing facility is only assigned to an episode where it is closely related to the inpatient surgical procedure. That is detailed in the Measure Information Form, Section A-3.

For the testing, the developer used a signal-to-noise analysis to evaluate reliability at the group practice level, the TIN level, and also the individual clinician level, so TIN-NPI level, using a split-sample method, calculated from a larger sample of episodes in 2018 and 2019, to get enough volume per TIN and TIN-NPI. So the minimum episodes were noted as ten for TIN and TIN-NPI.

The developer calculated Shrout-Fleiss intraclass correlation coefficients. The mean signal-to-noise reliability was 0.78 at the TIN level, and 0.72 for the TIN-NPI level. Reliability was slightly lower at the tenth and 25 deciles, 0.64 and 0.69, respectively, at the TIN level, 0.60 and 0.65 at the TIN-NPI level. And the developer also noted that they were higher at the

90th percentile, so 0.92 at the TIN, and 0.84 at the TIN-NPI.

Reliability at the practice size is also evaluated with the average reliability scores increasing from 0.71, which was one clinician, to 0.95, 21-plus clinicians at the TIN level. Pearson correlation and ICC coefficients between the split sample measure scores were 0.73 at the TIN level, and 0.67 at the TIN-NPI level.

The SMP did not raise any other major concerns, and passed the measure on reliability. And the staff also rated -- the staff preliminary rating was moderate. So I'll hand it back over to Kristine. Thank you.

Chair Martin Anderson: Thank you. So focusing first on reliability, any questions or comments from the committee?

Okay. Want to move to validity?

Ms. White: Okay. Thank you, Kristine. And I'm going to pull a comment out of the chat, as we keep going, just to get these on record. So this comment is coming from Risha. She provided a suggestion for the future.

She's seeing a base plus race model only would be very helpful. She's reviewing models, and she sees the base and duals and race models, but that it's hard to parse, since half of duals are of a minority group, so a lot of the variation due to race is likely being sopped up by the dual. Having a base plus race model only required for presentation would be helpful to elucidate this. And so thank you, Risha, for that feedback, and we will definitely take that back and --

Chair Martin Anderson: And LeeAnn, that's also consistent with some conversations we've had in the past, about the partial effects, et cetera. And I thought NQF was going to be doing a special project on this. Or have you already? I don't know, Matt, have you done that, yet?

Dr. Pickering: So with respect to social risk factors in risk adjustment, the quality measures, this project is still being finalized. But that is going to be walking through some step-by-step guidance that developers should consider, and some of the expectations related to how to consider social risk factors, as well as functional risk factors in risk-adjustment models. The suggestion, Risha, that you had mentioned, it's a good one, and maybe we'd take that to the developers, as well. We see in the chat that LeeAnn has summarized.

Related to this project that's ongoing, that should be finalized by the end of this year, and then from there, we would then reviewing those recommendations made from that project into our measure evaluation criteria and policies. So that would be something that would have to be the next phase of all of this work.

So all that to say is, it's still ongoing, and it's still going to be finalized by the end of this year, in which point we would then need to update our policies and criteria. So at this point, it's what the developer has submitted to you thus far, and their rationale, to consider and take into consideration for your voting.

Chair Martin Anderson: Yes. Yes. I was just thinking of all the comments we've gotten from Risha and Cheryl in the past, et cetera, et cetera, if you guys were sopping all those up -- for that deep dive, that'd be awesome.

Ms. White: Okay. Kristine, do you want me to go on to the validity testing and --

Chair Martin Anderson: Yes.

Ms. White: Okay, perfect. Thank you.

Chair Martin Anderson: Perfect.

Ms. White: So for validity, we have several subcomponents of this criterion. So we look at the testing exclusions, risk adjustment, meaningful differences, multiple data sources, and disparities.

So I'll start with the specifications that align with the measure intent for attribution. So the developer noted that this measure attributes lumbar spine fusion episodes to the clinician, so the TIN-NPI, that's billing the triggering procedure code.

At the clinician group level, an episode is attributed to the TIN if the TIN-NPI attributed an episode by billing the triggering procedure, and all episodes across the TINs-NPIs are aggregated. If the same episode is attributed to more than one NPI within a TIN, this episode is only attributed to the TIN one time, or once.

For costs, the cost approach, the developer noted that this measure uses Medicare standardized pricing. The methodology used to payment-standardize the Medicare claims is available for download, so they do provide that in the measure worksheet for your convenience and to access that.

For validity testing, the SMP did review the validity testing of the measure, and they passed the measure on validity. The developer conducted both face validity and empirical validity, so I'll start with the face validity.

The developer convened a clinical subcommittee, which included 22 members from relevant clinical experience in a Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels, Clinician Expert Workgroup of 13 members. Workgroup members -- so nine out of 13 members voted, which was a 69 percent response rate, and they agreed that the measure could accurately capture a clinician's risk-adjusted cost to Medicare for patients who receive lumbar spine fusions, with mean ratings of 3.9 or higher out of a scale of six for five base validity questions related to triggers, exclusions, service assignments, episode window identification, and risk adjustment variables. The mean response ratings was 4.5 on all five questions, or somewhat to moderately agree.

The developer was unable to obtain a mean rating on

the question, the score obtained from a non-emergent lumbar spine fusion measure as specified will provide an accurate reflection of cost per episode of care, and can be used to distinguish good and poor performance on cost effectiveness.

For empirical validity testing, the developer evaluated and examined correlation with the Medicare Spending per Beneficiary, which is MSPB, Hospital Measure, NQF 2158, which assesses the risk-adjusted cost to Medicare for services performed by hospitals and other healthcare providers during an MSPB hospital episode. Specifically, the developer analyzed the distribution of lumbar spine fusion measure scores, which is observed-expected cost ratios across the MSPB performance ratings.

Their empirical testing showed that the mean score cost -- or mean cost scores, which is observed over expected ratios, were highest for TINs with lowest performance on the MSPB hospital measure, low cost efficiency at 1.04, decreasing as performance ratings increased to 0.96 at performance ratings from five to ten, best cost efficiency, as expected. A similar result for the TIN-NPI, with mean cost score 1.04 for lowest performance rating, 0.96 at highest performance rating.

For clinical inclusions and exclusions, evidence to support the clinical logic, the developer excludes certain episodes -- patients with cancer, patients with an infection, patients that underwent a redo of a lumbar fusion -- to achieve fair comparison across providers. The developer also reports that the statistical results of the exclusions provide evidence that the excluded episodes are not comparable to the overall measure of population.

Moving on to risk adjustment, the developer included 122 risk factors in the overall risk model. The risk model was informed by covariates recommended by developer-convened expert panels and the CMS Hierarchical condition categories, HCC, as well as the demographic information from the Medicare

enrollment files. So age, race, disability, dual status. Information on income, education and unemployment were obtained from the Census American Community Survey's data.

The risk adjustment models formed separately for three measure subgroups, based on the level of fusion. So one-level lumbar spine fusion, two-level lumbar spine fusion, and the three-level lumbar spine fusion. The social risk factors were included after the base risk adjustment for clinical factors.

Stepwise regression was used to include sex and dual status and race, sex and dual status and income, education and unemployment, age against -- dual status, Agency for Healthcare Research and Quality, socioeconomic status index, sex plus dual status, plus income and education and unemployment, and race and sex, dual status, race, AHRQ SES and sex. So several combinations.

The reporter -- or developers report that the analyses found that the relationship between the various social risk factors tested and the measure cost scores were inconsistent across factors, and sometimes negative. The developer also reports that including these factors could introduce bias into the measure.

Many significant P-values indicate a social risk -- indicate social risk factors of predictive or resource use. However, analysis results suggested that adding social risk factors to the measure risk-adjustment model had minimal impact on the measure performance. It was largely redundant with current model predictions.

The developer also determined, using two methods -- this was determined using two methods -- analyzing differences in percentiles of observed to expected cost ratios, both with and without social risk factors in the model, and examining correlations between measure scores calculated with and without social risk factors. Both of these tests demonstrated a minimal impact on the measure's performance, even for providers at high and low extreme risk from

including social risk factors in the model.

So overall, the developer noted that the overall R-squared for the measure was 0.516, adjusted R-squared of 0.513. And the average observed-to-expected cost was generally close to one, 0.9992101, across risk deciles. And the average observed-to-expected cost ratios for all risk deciles are close to one. For meaningful differences, at the TIN level, the standard deviation is .09, and at the TIN-NPI level, the standard deviation is .10. Scores were not influenced by region or number of episodes performed.

This measure uses Medicare administration claims only, so there is no multiple data sources. The developer did not provide performance data distributions by subpopulations under the disparities. And the staff preliminary rating for validity was moderate. I'll hand it over to you, Kristine.

Chair Martin Anderson: Thank you. Just to get us started here, there's a couple of things that are a little bit different about this measure. One is, it looks like a lot of the testing was done in the three sort of subgroups. Right? And then, I presume, overall. I wondered if you could verify that, because that was -- in the data, sometimes I was getting a little lost in the way that you broke out the testing.

And then, related to that, also, for the measure developers, it looks like it was a bit harder to decide what services to include in this measure, a little bit more variability in consensus around which services belong in the episode. Can you comment on both of those?

Ms. Lam: Sure. So for the first item, for subgroups, so the way that the subgroups work is that basically, we stratify all the episodes into these three mutually exclusive subgroups, which represent the three different levels of procedures. And so we run the risk adjustment models separately within each of these subgroups. So that's why the risk adjustment testing shows the results at the subgroup level. Then, to turn

it into a score, we rolled it up to the provider level. But for the purposes of risk-adjustment testing, we showed the results for each model as they're run within the subgroup.

Chair Martin Anderson: Okay.

Ms. Lam: And then for service assignment -- so we went through a similar process as the other measures. And the discussions about which services to include -- we do use a voting process, where we work with the clinical expert panels, where we discuss specific types of services, the timeframe for including them.

And so I'm just checking to see if we have any particular areas where it seemed like it was more difficult to reach consensus. I think, from memory and from just looking at the codes list, they generally fall within the buckets that reflect the areas of care. So it's things that are common to the other measures, like complications, imaging, and readmissions, post-acute care use.

Chair Martin Anderson: Thank you. I had seen in the Scientific Methods report that they were saying that there were some areas where, you know, it looked like it was more divided, in terms of, you know -- when you look at the Scientific Methods report in total, it's just an area that they seemed a little less comfortable with exclusions and service categories than with the other two, but ultimately, did give it a moderate rating. Thank you. Anyone else have comments about validity, or questions?

Okay. I think we're ready to go on to usability and use. No -- is it feasibility? Yes. Great.

Ms. White: Feasibility.

Chair Martin Anderson: Feasibility.

Ms. White: Wonderful. So this, feasibility, the measure developer indicates that all the data elements for this measure are in defined fields, in a

combination of electronic forms, and uses Medicare administrative claims data, and that there are no associated fees with the use of this measure. The staff preliminary rating for feasibility is high. And that is all I have for feasibility. I do not see any concerns that were brought up through the pre-evaluation survey.

Chair Martin Anderson: Thank you. Any comments from the committee?

Okay. Now onto usability and use.

Ms. White: Wonderful. So we'll start with use first. The measure developer indicates that this measure is publicly reported and used in accountability programs. The measure is currently used in the Quality Payment Program, Merit-based Incentive Program System, so MIPS, and as specified in the Calendar Year 2020 Physician Fee Schedule Final Rule, this measure has been implemented as part of the MIPS program, beginning in the 2020 MIPS performance year.

For the feedback on the measure by those being measured or others, the developer indicates that this measure -- the lumbar spine fusion field test reports were provided to a sample of eligible clinician groups and clinicians.

Each report did include information for the lumbar spine fusion measure if the clinician or clinician group was attributed to ten or more episodes. All stakeholders, including those who did not receive a field test report, could review a mock field test report that was posted on the CMS website.

During the field testing, the developer conducted education and outreach activities, including the national webinars, office hours with specialty societies, and Help Desk support. The developer also sought feedback on the reports and measure specifications through online surveys, with an option to attach a comment letter.

After completing field testing, the developer compiled the feedback provided through the survey and comment letters into a measure-specific report, which was then provided to the Lumbar Spine Fusion Clinical Expert Workgroup, along with the empirical analysis, to inform their discussion and evaluation of any refinements needed to ensure that the measure is capturing what it's intended to capture.

Stakeholders provided cross-cutting feedback on risk adjustment variables. For example, the cognitive and functional status, academic medical centers, and socioeconomic status. They also provided feedback on attribution methodology, episode windows and assigned services, and alignment with cost and quality.

For additional feedback, the developer noted that the lumbar spine fusion measure was implemented in MIPS after going through the pre-rulemaking process and notice-and-commenting rulemaking. The measure was submitted to and included in the 208 Measures Under Consideration list.

It was then considered by a National Quality Forums Measure Applications Partnership Clinician Workgroup and coordinating committee in December of 2018, and January of 2019, respectively. The MAP voted to conditionally support this measure for rulemaking, conditional on submission to the NQF review and endorsement process.

The MAP noted that CMS and the Cost and Efficiency Standing Committee should continue to evaluate their risk-adjustment model of this measure, and consider whether there is need for -- a need to account for social risk factors in this model.

The MAP also noted that the review of the measure should ensure an appropriate attribution methodology, and that the measure adequately considers the issue of small numbers. The MAP additionally notes that the cost measures should continue surveillance for unintended consequences such as stinting of care and reduced quality of care,

and that cost measures should be paired with balancing measures -- so for example, quality, efficiency, access, and appropriate use measures -- as a way to safeguard against these issues.

Lastly, the MAP recognized a need for continuous feedback and testing of measures as they are implemented, and agreed to provide greater education on these measures, as well as for greater transparency of the measure specifications and testing results.

The staff preliminary rating for use is pass, and there were no concerns that were brought forward in the pre-evaluation comments by the standing committee. And I do see, Kristine, that Risha has her hand raised, as well.

Chair Martin Anderson: Great. Go ahead, Risha.

Member Gidwani: Hi, thanks. You know, this question I have is not just specific to this measure, but actually across all three measures submitted today. In all sections on unintended consequences, the developers noted not applicable.

And that was surprising to me. You know, even just thinking about disparities, going back to the previous measure for CABG, we know that Black patients are less likely to receive recommended medical care for CABG, more likely to have complications.

Whenever we're looking at cost measures in isolation, without some explicit tie to quality or formal tie to quality, which we don't have here -- even though it might be in a global set of measures, they're not explicitly together -- the incentive is always going to be with withhold treatment or to undertreat. And so it was very surprising to me to see the developers note no unintended consequences, when we know about racial disparities and undertreatment.

Can the developers speak to this, and sort of help me understand why you felt that it wasn't applicable to discuss positive -- I'm sorry, potential negative

externalities to disparities?

Mr. Nagavarapu: Sure. I could start in on that, and Joyce, Heather, Rose, if you want to jump in, feel free. Yeah, I think it's a really important question. The main reason we didn't talk through the specifics of unintended consequences in the form is that there are kind of two real countervailing features of the measures, and how they're going to be used in the programs, that can help address the type of concerns that you're talking about.

The first is something that is just an important feature of the measure construction, in that, as we mentioned, the cost drivers in the measure construction have to do, in large part, with adverse outcomes. And so if you get sort of a classic care stinting going on, where there's less care going on, either around the initial procedure, or in cases where post-acute care is needed, the fact that that could lead to a complication that's captured in the measure can sway performance in a really big way.

And so, like, the example that Joyce gave earlier was, with the unplanned readmission algorithm, if you look at the difference in observe cost between episodes with and without an unplanned readmission, it's $53,000 versus $37,000. So I guess almost a 50 percent increase off of the base.

The other related fact that we didn't talk about, but it was really compelling when I saw it, was if you look at the first decile of the performance score for the cost measure, to the tenth decile of the performance score -- so lowest cost to highest cost -- and then look at the mean unplanned provider readmission rate, from also lowest cost to highest cost, lowest cost is at about 8.5 percent. The highest cost is at 34 percent.

And so it really is the case that these low-cost providers don't seem to be providers that are stinting on care in a way that's leading to worse quality outcomes, and their unplanned readmission rates are actually about a third of the rates of the highest

decile.

The other aspect of this was the -- sort of the program function of the measures. As Joyce mentioned, the vision for MIPS at CMS is being able to take these quality measures that are available for this type of cost measure, and use both the quality and cost measures in the construction of MIPS composite scores.

And so the hope is that, you know, there's going to be other quality outcomes besides unplanned readmissions that may not be caught in the cost measure. So that could be, like, functional status, things like that, that are really important quality outcomes. And so the hope is that there's the quality measures in the quality category of MIPS that are weighed alongside the cost measures that could help with that.

But it's definitely an important concern, and it's something that, like, our monitoring going forward is trying to get a sense of for each measure, at least to the extent that we can tease that out in claims data.

Member Gidwani: Okay. So you're saying that, pretty much, undertreatment will generally result in costly adverse events that will be captured within the 120-day, the 90-day post period, and so that's sort of your safety for ensuring that there's no undertreatment. Is that correct?

Mr. Nagavarapu: Exactly. And, you know, that argument helps for a lot of cases. I think there's a set of concerns that you might have, where it doesn't help for it, necessarily. Like, functional status is something that is not captured in claims data. Sort of, how easily can you go to the grocery store? And that's something where the hope is to rely on the type of quality measures in MIPS that Joyce mentioned, as well as ones that would be developed in the future.

Member Gidwani: Mm-hm. Okay. Is there -- I'm going through the Excels, and I'm not seeing this, but

maybe it's somewhere else. But do you guys have information that tells the review panel the contribution of various types of events towards higher costs? So for example, how much of the higher-cost providers are higher costs due to adverse events, due to readmissions, due to choosing higher-cost labs, due to choosing higher-cost imaging?

That would sort of be, I think, helpful in helping understand, I think, the risk of undertreatment, because if a lot of the costs are coming from adverse events, then yes, you're right, undertreatment would be captured in this 90-day postoperative period, because there would be adverse events.

But you know, if the costs are coming from the presence or absence of, you know, giving someone an x-ray versus an MRI, or giving someone a CBC versus a different type of lab, you know, not only -- may not move the needle quite as much, but that may result in a risk of undertreatment that wouldn't necessarily be always captured in the 90-day post period.

Mr. Nagavarapu: Yeah, absolutely. I mean, I think in future NQF submissions, we could summarize that breakdown. It's a breakdown where, unfortunately, like, we have that breakdown at the provider level, for the field testing reports.

So, like, when we send out the field testing reports, there's different categories for the different sort of big-ticket, like, types of complications that you can have, as well as things like imaging, that you mentioned. We unfortunately don't have an aggregated version of that on hand right now, but that's certainly a thing we can certainly include. I can see why it'd be useful to you.

Member Gidwani: Okay. Thank you.

Chair Martin Anderson: Okay. Emma, is your question also on unintended consequence, or something related to usability, or on use? Because I might want to have LeeAnn do her intro, if it's on

usability.

Member Hoo: You know, it's more of a general comment, insofar as -- I know we can only score, you know, what's in front of us. And when PBGH managed a Centers of Excellence program in the space of spinal, you know, fusion surgery, one of the huge issues was inappropriate surgeries that weren't going to generate an improved outcome to begin with.

And I think, you know, one of the things I find challenging, just in this exercise and discussion, is the ability to, you know, filter for those cases that received this procedure that aren't benefiting at all. And you know, they may be lower risk, or they may range in the cost profile, but you know, it's hard to gauge what that would look like in an environment where there is more of a Centers of Excellence program that screens out the avoidable services, versus one that is less discriminate on the cases.

Chair Martin Anderson: Certainly this process presumes appropriateness. Right? And I get what you're saying, that that, in and of itself, may be the real issue in this area. Okay. LeeAnn, why don't you cover usability? I have one more comment, but I think it fits better in usability.

Ms. White: Perfect. So for usability, the developer did provide improvement results, a distribution of performance scores, for the clinician group level and the individual clinician level, that attributed ten or more lumbar spine fusion episodes from January 1st of 2019 through December 31st of 2019. For the group level, the mean was 1.01, with an interquartile range of 0.10. And for the individual level, the mean score was 1, with an interquartile range of 0.11.

The developer further notes that there were no unintended consequences to individuals or populations identified during the development and testing of the measure, and no potential harms were identified.

The staff did have a preliminary rating for usability of

moderate. There were some -- there were no -- we did receive some feedback with the pre-evaluation summary. One standing committee member noted, it's not clear by the information provided whether this measure is currently publicly reported or not.

And then there is a potential of harm in that the accountability for cost in the short term maybe influence decision-making that may have an impact beyond the episode of measurement. And that concern was raised in the benefits versus harm portion of usability. But I'll hand it over to you, Kristine.

Chair Martin Anderson: Thank you. And I'll just start here with something that I'm not really sure how to deal with. But it bothers me, both in unintended consequences and just in this topic, broadly.

The use of opioids pre-surgery has been linked to, you know, the quality of the surgical outcome, and if there needs to be a second surgery, et cetera -- which likely would not happen in 90 days, and would create a whole new episode -- then opioid use and opioid problems with use tend to go way up.

And so, you know, what I'm struggling with is the lack of drug data in this particular condition, given what we know about, you know, what happens with opioids both pre- and post-surgery. I don't know if that is something that has come up in your clinical committees, but measuring lumbar spine surgery without considering drugs feels, to me, to be incomplete.

Ms. Litvinoff: Yeah, hi, this is Heather from the Acumen team. Certainly, that is something that many of our committees have discussed. We do have medications included in the service assignment here. But certainly, I think it's really important also to rely on quality measures that also focus on opioid use, as those are obviously, you know, looking very specifically at practices around the prescription and use of those medications. Hopefully that's helpful.

Chair Martin Anderson: Yeah, I don't think it's one we can solve. Right? But it just feels like the windows need to be broader if you're going to consider that, and then also the, you know -- and so did you have -- do you have a risk-adjusted for opioid use prior to surgery?

Ms. Litvinoff: Let me just -- I do not know that we included that in this measure. But let me see if other members of the team have other comments on that.

Chair Martin Anderson: Yeah, I don't think it's solvable. It's just -- I just lay it out there as something that bothers me.

Member Gidwani: While they're looking, Kristine, yeah, I agree with you. I think, if I recall correctly, no prescription medicines were included in these measures, and I wasn't sure if that was an artifact of the patients potentially not being enrolled in Part D, and therefore not having complete drug data on all of the patients. Was that the reason for exclusion of prescription drugs from all of these measures?

Ms. Lam: Yes. So the Part D drug costs, that's something where, at the time of development, we weren't able to include in the measure, because there weren't any standardized Part D costs available. And now, there actually are standardized Part D costs. So they can be considered for inclusion in cost measures.

And two of the -- three of the cost measures in MIPS for 2022 actually do include Part D costs. So that's the chronic condition measure for diabetes, asthma, COPD, and the inpatient measure for sepsis.

And we got some great feedback from our tech, who gives us guidance on overarching considerations, and they took this under consideration, that workgroups should think about, when thinking about whether or not to include Part D costs, because it's not always going to be relevant to every type of care being assessed.

And so some of the considerations include, what share of episodes include Part D costs? How much does it make up the assessment of clinician performance? So you could think about, for a diabetes measure, not including insulin could lead to a less accurate reflection of performance.

The way that we account for the fact that not everyone is enrolled in Part D is, we subgroup by Part D enrollment, so that costs are only compared amongst bennies with Part D or amongst bennies without Part D. and so --

Mr. Nagavarapu: And -- oh, sorry, go ahead, Joyce.

Ms. Lam: Oh, no, you go.

Mr. Nagavarapu: Oh, and so I was just going to say that, at the tie of development of these measures, as Joyce mentioned, there are -- Part D standardized costs were not yet available, and the workgroups were very concerned about drug price variation that clinicians couldn't control.

And so specifically when the workgroups chose the measures to develop -- so, like, the three measures that we're discussing today, as well as other measures that were developed around the same time -- they specifically chose measures that they felt good about developing, knowing, going in, that they wouldn't be able to include Part D costs.

And so they were aware of that, and kind of made a choice that this episode group could be a functional and useful episode group, even without Part D cost, but that once Part D standardized costs did become available, they might be able to revisit that decision and see whether in the next round of comprehensive reevaluation, that they'd like to adjust that decision.

And for the opioid question, I think it's a huge question. You know, I don't think there's anything in terms of risk adjustment specifically here for it. I think there's a lot of issues that I'd be interested to hear all of your thoughts on, as well as other

stakeholders down the road, just in terms of, for instance, even access to the new Medicare opioid prevention programs -- the treatment programs, I mean -- and how that varies across geography and so on. And is there a rationale for potentially considering that type of variation in risk adjustment down the road?

And so if, you know, beyond the HCC, for drug dependence, that is in the risk adjusted model, we don't have something specifically for opioids, but we're certainly open to feedback on that. I think it's a really complicated issue here, and we'd be happy to take any thoughts you have on that back to CMS.

Chair Martin Anderson: Yeah, I just think that's something that you should look at, particularly in this particular area. Right? If you look at which surgeries are associated with, you know, the highest risks for opioid dependence, and/or people who have already been on opioids prior to surgery, you know, there are some abdominal -- excuse me -- and then also, obviously, some, like, chronic pain related surgeries. And I think if we just measure pain management, and we don't consider, you know, drugs, we're going to miss something big. So for future research.

Mr. Nagavarapu: Yeah, no, thanks very much. Yeah, that's, yeah, beyond the drug dependence HCC and the risk adjustment, we're not accounting for specifics, and so that's certainly something we could keep in mind in the future.

Chair Martin Anderson: Okay. Any other committee members have any questions or comments on use and usability?

I really don't want to let you go two and a half hours early. It's just, you know, awful. Just kidding. Okay. I think -- back to you, LeeAnn. I think we've completed our review of the measures. I guess a reminder to vote on SurveyMonkey before you forget what you would vote. And you want to go through the next steps, LeeAnn?

Next Steps

Ms. White: I would love to. Yes, we are ahead of schedule, which is great. We can give you some time back to your Tuesday. We do need to do a few more items on our agenda. So I'm going to just pause and have our team member pull up our slide deck real quickly here.

Our next item on the agenda is our NQF member and public commenting time period. This will -- we will open the lines and ask NQF members and public if they'd like to comment on the measures that were under review today, or any of the discussions that were had. So I'm going to pause here for a few moments to open up the lines for that.

NQF Member and Public Comment

Ms. White: Hearing no comments come through, I will hand it over to Tristan Wind, who will take us through our remaining activities and upcoming timelines for this project. So Tristan, it's all yours.

Mr. Wind: Thank you, LeeAnn. Next slide, please. Thank you for attending today's call. Following the conclusion of this meeting, NQF staff will prepare a draft report, specifically noting the standing committee's discussion and recommendation, which will be released for a 30-day member and public comment period.

Staff will incorporate the comments received into a comment brief. That'll be shared with the standing committee and developers. These comments will then be discussed in the post-comment call, in which staff will incorporate comments and response to comments to prepare for the CSAC meeting, in which CSAC meets to endorse measures. Lastly, there is an opportunity for the public to appeal the endorsement decision. Next slide, please.

The draft report comment period is from August 15th to September 13th. The dates for the post-comment meeting, CSAC review, and appeals period has not

yet been finalized. staff will communicate those dates accordingly. Next slide, please.

And lastly, here's the project contact information, if you happen to have any questions or need assistance following the conclusion of this call. Additionally, the project page and committee SharePoint site are listed to access final project deliverables. I will now turn back to LeeAnn for outstanding questions and closing remarks.

Ms. White: Thank you, Tristan. So just to bring forward what Kristine was mentioning, as well, with the SurveyMonkey -- so definitely will get that out to the entire standing committee here shortly.

We will convert today's call into a recording, and we will be sending the recording link, which is a video, and the SurveyMonkey link to all standing committee members who were not able to join the call, as well, today, and give them that opportunity to review the measure discussions and then vote on the criteria. We will put the deadline for voting within the email. And so that will be occurring either late this week or next week. You should see that come through email.

Also, I just want to pause a moment to see if anyone has any questions about the call today, or the measure discussions, or the next steps.

Okay. Well, I want to thank everyone for attending today. It was a really robust day. Great conversations. This was my first Spring 2022 Cost and Efficiency Measure Eval meeting, and I definitely appreciate everyone's patience.

I, again, thank you for attending today's call and for being present and engaging. I learned a lot from everyone today. And thank you so much for the developers for being on the call to present their measures. We greatly appreciate your attendance and your participation leading up to the meeting. So great big thank you to the standing committee members, the developers.

I also would like to thank our esteemed co-chairs, Kristine and Sunny. Thank you so much for leading us through the measure evaluations today. We definitely appreciate your commitment to our project team and the standing committee.

I'll pause to hand it over to Kristine and Sunny, so that they can say their closing remarks to the standing committee, as well, and attendees.

Chair Martin Anderson: Go ahead, Sunny.

Chair Jhamnani: Thank you, guys, for being present today. Thank you for taking out the time to review these measures, and for your thoughtful comments. There's a lot of work that needs to be done. I think over the past two years, we've pushed the bar higher and higher, and that is a testament to the work and the expertise that you bring here to the table.

And I think, as we continue this work, I think Acumen will continue to raise the bar and bring a more robust measure for us to battle with next time, and CMS. Thank you again for your time.

Chair Martin Anderson: Thank you all. Well said, Sunny. Thanks, and enjoy your afternoon. Some found time.

Ms. White: Thank you. Thank you. Thank you so much.

Chair Martin Anderson: Bye. Thanks to --

Ms. White: Take care, everyone. Be safe.

Chair Martin Anderson: Thanks to the NQF staff and the developers. Bye.

Ms. White: Yes. Thank you. Bye.

Adjourn

(Whereupon, the above-entitled matter went off the record at 2:31 p.m.)