NATIONAL QUALITY FORUM

+ + + + +

CARDIOVASCULAR STANDING COMMITTEE

+ + + + +

CARDIOVASCULAR MEASURE ENDORSEMENT PROJECT 2015

+ + + + +

THURSDAY
SEPTEMBER 10, 2015

+ + + + +

The Cardiovascular Standing Committee met at the National Quality Forum, 9th Floor Conference Room, 1030 15th Street, N.W., Washington, D.C., at 8:30 a.m., Mary George and Thomas Kottke, Co-Chairs, presiding.

PRESENT:

MARY GEORGE, MD, MSPH, FACS, FAHA, Co-Chair
Senior Medical Officer, Centers for Disease Control and Prevention, Division for Heart Disease and Stroke Prevention

THOMAS KOTTKE, MD, MSPH, Co-Chair, Medical Director for Population Health, Consulting Cardiologist, HealthPartners

SANA AL-KHATIB, MD, MHS, Associate Professor of Medicine, Duke University Medical Center

LINDA BRIGGS, DNP, Assistant Professor, George Washington University, School of Nursing

LESLIE CHO, MD, Section Head, Preventive Cardiology and Rehabilitation, Cleveland Clinic

JOSEPH CLEVELAND, MD, Professor of Cardiothoracic Surgery & Surgical Director for Adult Cardiac Transplantation/Mechanical Cardiac Assist Devices, University of Colorado Denver

MICHAEL CROUCH, MD, MSPH, FAAFP, Research Director and Quality Improvement Program Director, Memorial Family Medicine Residency Program and Associate Clinical Professor of Family Medicine, Texas A & M University School of Medicine

ELIZABETH DELONG, PhD, Professor and Chair, Department of Biostatistics and Bioinformatics, Duke University Medical Center

ELLEN HILLEGASS, PT, EdD, CCS, FAACVPR, FAPTA American Physical Therapy Association

JUDD HOLLANDER, MD, FACEP, Associate Dean, Strategic Health Initiatives, Sidney Kimmel Medical College, Professor, Vice Chair of Finance and Healthcare Enterprises, Department of Emergency Medicine, Thomas Jefferson University

THOMAS JAMES, MD, Chief Medical Officer, Baptist Health Plan and Baptist Health Community Care (via telephone)

JOEL MARRS, Pharm.D., FNLA, BCPS (AQ Cardiology),CLS Assistant Professor, Department of Clinical Pharmacy, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of Colorado Anschutz Medical Campus, American Society of Health-System Pharmacists

GERARD R. MARTIN, MD, Senior Vice President, HLK, Medical Director, Global Services, Children's National Health System

KRISTI MITCHELL, MPH, Senior Vice President, Avalere Health, LLC

GEORGE PHILIPPIDES, MD, Chief of Cardiology, Newton-Wellesly Hospital

NICHOLAS RUGGIERO, MD FACP FACC FSCAI FSVM FCPP, Director of Structural Heart Disease and Non-Coronary Interventions, Thomas Jefferson University Hospital

JASON SPANGLER, MD, MPH, FACPM, Executive Director, Medical Policy, Amgen, Inc.

MLADEN VIDOVICH, MD, Chief of Cardiology, Jesse Brown VA Medical Center

NQF STAFF:

HELEN BURSTIN, MD, MPH, Chief Scientific Officer
JASON GOLDWATER, MA, MPA, Senior Director
DONNA HERRING, MPH, Project Analyst
LAURA IBRAGIMOVA, MPH, Project Analyst
KAREN JOHNSON, MS, PhD(c), Senior Director
MELISSA MARINELARENA, RN, MPA, Senior Director
LESLIE VICALE, MPH, Project Manager
ASHLIE WILBON, MS, MPH, FNP-C, Managing Director

C O N T E N T S

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

8:36 a.m.

MS. VICALE: Good morning, everyone. We'd like to start Day 2 of the Cardiovascular Measure Endorsement Meeting. I'd like to ask our co-chairs, Mary George and Tom Kottke, to welcome everyone for the day.

CO-CHAIR KOTTKE: Well, welcome. I'm not going to say much more, we have a full agenda. Thank you all for showing up on time and I'll turn it over to Mary.

CO-CHAIR GEORGE: Yes. Thank you. And I just have one comment. Yesterday, we were doing a really good job of turning our microphones on and off, but sometimes it was a little bit hard to hear because you really weren't speaking into the microphone. So if you can remember to speak into the microphone, that would really help. Thank you.

MS. VICALE: Thanks, Mary and Tom. Next slide, please, yes. So before we get started, we wanted to just remind you all where the restrooms

are, if you exit the main conference area past the elevators on the right. Today, our breaks are scheduled for 10:15 for 15 minutes and for 12:30, where we'll have lunch. Again, you'll see here on this screen the wi-fi network and the password to join that. And as a reminder, we do appreciate if you mute your cell phones during the meeting to reduce any noise. And we ask folks on the phone to keep your lines muted to reduce any background noise. Okay.

So just a few items that we wanted to recap from yesterday about the ground rules. I won't go through all of them, really just some of the more important ones. So during the discussions, we ask that you base the evaluation and recommendations on the measure evaluation criteria and guidance. And you are very well aware of this, however we did want to just mention it again. Again, please try to stay in the room for all of the meeting. And keep your comments concise and focused and clear. And please do indicate your agreement without repeating what's already been

said.

And we are mindful of the time today. We will again be using the time cards that we had used yesterday. So if we hold up the yellow card, that means we have five minutes left for the measure. And if we hold up the red card, that means we have two minutes left to review the measure. And we are looking at about 20 to 30 minutes per measure for the day. We do, like Mary had said, ask you to please speak clearly into the microphones. And for any measures that you may need to recuse yourself from, we just ask you to do that before the discussion of the measure.

Again, I'm not going to go through the entire slide, however, we did want to remind you for the criteria, the Importance to Measure and Report, that's a must pass as well as the Scientific Acceptability of Measure Properties, that's also a must pass. Okay. Next slide. So at this time, I'm going to turn it over to Jason Goldwater who's going to highlight a few notes for the eMeasure evaluation that we have coming up today. Thanks.

MR. GOLDWATER: Thank you, Leslie, and good morning, everyone again. So, I'm sure this will be the highlight of your day, which is to learn about how we evaluate eMeasures. Clearly what everyone came to this meeting this morning wanting to learn. So, I'll do my best to be as entertaining as possible in the next five to ten minutes. As all of you know, eMeasures have been around for some time, this is certainly not a brand new concept. eMeasures have really started in the mid part of 2000s with CMS initiatives and have grown since that point in time. Some of the projects that we have had over the last few months, there has been a steady increase in eMeasures, particularly around this topic area.

And what we wanted to do today was to just talk about how we evaluate eMeasures independent of how the measure is evaluated normally and things to look for when you consider eMeasures today and in the future. And really sort of how we look at eMeasures generally speaking. eMeasures are considered separate measures from a

traditional measure because it's generally based on a data source. Measures in the past, particularly around cardiovascular, a lot of you know that when quality measurement started two decades ago some of the first measures that came out were related to cardiovascular disease. And those measures were derived from generally claims-based data. And those paper measures have been around for some time.

eMeasures, and there's two types of eMeasures. There could be brand new de novo measures, and we heard about one yesterday that was being considered for the trial-use program. Or respecified measures, which is basically taking a paper measure that's been around for some time and may actually have an NQF number and has been endorsed previously that is being respecified into an eMeasure.

And what we mean by respecification is, it's utilizing a new data source to populate and report out on the measure, which could either be from a registry-based system or from an electronic

health record. It is formatted very differently because it has to be transferred from one system to another, and I'll talk about those formats in a moment. And it evolves, obviously, with the science as well because as structure data is included within the EHR registry, the measure can evolve along with it.

We provide a technical review of the measure, clearly because of NQF's position, we don't obviously say whether we agree with the measure or endorse the measure or pass any judgment on the measure. We only do essentially a technical review, which involves several parts. The first is to make sure that the measure is formatted correctly. An Electronic Clinical Quality Measure has to be specified in what was known as the Health Quality Measures Format, or HQMF for short. Without delving into geek-speak as I said yesterday and as Helen knows I'm very fond of --

DR. BURSTIN: Very.

MR. GOLDWATER: Thank you, Helen. The HQMF is really an extensible markup language. And

for those of you that do not know what that is, that is really how you connect to the internet. It's how you interact with the internet, it's how the internet transfers information from one system to another. We have a very defined format that an Electronic Clinical Quality Measure has to conform to.

It has to map elements of the quality data model, which means certain categories have to be filled, such as if it is an outcome measure, diagnosis has to be populated in one certain fashion, procedures, laboratory codes, medications, et cetera. It has to have value sets. And value sets are what are sort of known as the building blocks of measures. They basically represent a clinical concept. So bypass graft or bypass surgery would be a value set. And those value sets are encoded in a particular vocabulary or standard and measure Developers, as they will tell you today choose the value sets that best correspond with the intent of the measure.

When we do our assessment, we ensure the

fact that the measure is formatted correctly, that it has the appropriate elements of the quality data model, that it has value sets, which are actually curated and maintained by the National Library of Medicine, and those value sets are published in the Value Set Authority Center. We make sure that those value sets are published, they're not draft or they're not proposed, they are published and can be used by any measure developer and they are nationally recognized. And then we check for the feasibility of the measure as well. A measure developer has to send a feasibility scorecard and they have to provide justification as to why they are giving the scores that they are given.

The eMeasures are expected to meet the same criteria as all measures with some specific applications. The first, as I told you this yesterday, they have to test for reliability and validity within an electronic health records system. The old criteria used to be three or more. That proved to be challenging for some. So we switched that to more than one, so essentially two.

So they have to choose two EHRs.

And as I was explaining yesterday, there are vendors that have similar record systems, but they are considered two separate EHRs. So the Epic system which is used for in-patient and the Epic system that is used for ambulatory care, although it is the same vendor and in many ways it's the same format, they are considered two separate systems. So if a measure developer were to test in-patient and out-patient using that, that would meet our criteria.

The feasibility assessment, in addition to how we normally assess feasibility, we also look to address that the data elements are correct and that the measure logic is calculating correctly. That we can derive from the testing data that is provided by the developer or in some cases, although this will not be the case today, that they can do a simulated test using tools that are provided by the MITRE corporation and we can evaluate the logic that way. The case for today, they will have actually tested in actual EHRs or

registry systems and will be able to tell us through the results whether the data elements and measure logic were calculated correctly. Anything else? Sure.

MEMBER DELONG: Do you evaluate that or do we evaluate how well it does in more than one system?

MR. GOLDWATER: So, we evaluate whether they filled out the feasibility assessment. We examine the scores that they gave. And then we make sure that the justification is there. We do not assess whether it's adequate justification, that is something you need to be doing.

We basically -- so for example, if they scored something and had a scorecard and it was all threes, which means it's the highest score they can give, and they provide no justification for that, we have to send it back to the developer and say, you need to provide some reason for why you gave these scores because if we present this to a committee and they look at a feasibility scorecard and all they see are threes without any

justification, it is very difficult for you to make an honest assessment as to whether the measure is feasible or not.

MEMBER DELONG: So for their justification, it would seem that they would supply data --

MR. GOLDWATER: Correct.

MEMBER DELONG: -- not just text saying, well this looked good to us.

MR. GOLDWATER: They would provide data from the testing and then they would provide summaries of the results of that testing, which would indicate whether the data was feasible or not. Yes, ma'am?

MEMBER MITCHELL: Would you mind taking a moment and walking us through the BONNIE output pre-testing? What is that?

MR. GOLDWATER: Sure. So I'm going to get to that in this slide. So that was rather serendipitous, thank you. So, what we're looking at today are what we call re-tooled measures. And I should say that we're not really fond of the word

re-tooled and we really are not trying to move to respecified, because that's essentially what they're doing. They're not necessarily re-tooling the measure, they're just respecifying it to be electronic. And, again, they basically take an existing measure and they respecify it to be an Electronic Clinical Quality measure.

Current NQF policy considers eMeasures as a separate measure. However, both are used in federal programs, such as PQRS, Meaningful Use, and are using the same number. The eMeasure and the claims registry measure are considered separately in this particular evaluation because they're using two different data sources. The measure that is using registry data and the measure that is using electronic health record data are considered two separate measures and have to be evaluated separately.

The BONNIE testing tool, so let me just explain briefly what that tool is and then when we accept it. And I'm not sure that's the case today, because I don't think these are legacy measures.

I don't believe they are from what I remember. The BONNIE testing tool was created by the MITRE Corporation and it was really designed for developers to test measures before they implemented them within their electronic health record.

And what you are able to do with BONNIE is create what's known as a synthesized data bank of patients. So you can basically create an N of 50 of patients and you can determine the characteristics of those patients, you can determine the demographics of those patients, based on the measure that you're testing. You can then run the measure against that test data bank to see if the measure logic is calculating correctly. So is it excluding the individuals that it needs to exclude? Is it including the appropriate individuals in the numerator and the denominator? Is it making the appropriate exceptions as needed?

It is just designed to make sure that the logic is calculating correctly and that the

measure's populating as it should. It is only a simulated test data bank though, it's not live patients. It's really to give us an idea that the way the measure is constructed was constructed correctly. And the BONNIE testing data is to provide that input. It was designed initially, of course, to allow vendors to test measures before they implemented them into their systems. But now it's been expanded to allow developers to test measures before they actually go through either live testing or in some particular cases, they can present BONNIE output here.

Now, where do we accept the output for BONNIE? There's only two situations in which we do that. The first is, yesterday when they do a trial-use measure. The reason a measure gets put into the trial-use program is because they are unable to do testing. But they can simulate testing through the BONNIE tool and make sure that the logic is calculating correctly, the numerator and denominator are being populated correctly, and the right exemptions and exclusions are being made.

So, once that's done, they can then present those results here, not to say that if the measure is implemented in the trial-use program it will function appropriately. We don't know that until the measure is out in the field and being used. What we can say through the results of that testing are that the measure logic is calculating correctly so we know when it's implemented in the field, at least from the structural standpoint, the measure is working correctly. Liz first and then Judd.

MEMBER DELONG: So when they do that test using the BONNIE system, if it's a re-tooled measure, do they provide a comparison between what they got from the BONNIE application versus the original application?

MR. GOLDWATER: Yes. So one of the things that they should do when they're doing a respecified measure is they need to provide the output from the initial measure as well as what they got from the BONNIE tool, correct. Yes, sir?

MEMBER HOLLANDER: So if there's a

conflict between what I'll call real data, with a little cynicism, and the BONNIE output, the real data should win, right?

MR. GOLDWATER: Yes, absolutely.

MEMBER HOLLANDER: Okay.

MR. GOLDWATER: Right. We would not use BONNIE as the basis for an endorsed measure. It's only to -- and that's only in the example of testing to make sure the logic is calculating. Yes, ma'am?

MEMBER MITCHELL: Just to clarify, the comparison is between the original specified output of the measure --

MR. GOLDWATER: Correct.

MEMBER MITCHELL: -- against the BONNIE. But then there's also the testing in the live EHR platform.

MR. GOLDWATER: Right.

MEMBER MITCHELL: So there's three buckets of data that need to be relatively aligned so that you feel confident in what you're seeing or no?

MR. GOLDWATER: No. So, BONNIE can only

be used, in our policy, it can only be used in two separate occasions. One is with trial-use. The other is in what we're considering legacy measures. And this is basically a stop-gap solution. So a legacy measure is a measure that is actually being used in a federal program. So it's being used in PQRS or it's being used for public reporting or IPR or whatever it may be, and they're respecifying that into an eMeasure. That has also proven to be difficult to test.

So a solution that was provided for just those measures -- now, keep in mind, those are already endorsed NQF measures that are already being used in a national program and are being respecified into electronic measures. If that is the case and only if that's the case, they can use BONNIE and present the output to us. Any other -- so if they create a de novo ECQM, they cannot use BONNIE. If they respecify a measure that's not in a national program, they cannot use BONNIE. They have to use the actual testing live data that has to be presented to you. They cannot use BONNIE as

a substitute for that.

And once we get to the point where they can actually test these legacy measures, eventually we'll sunset that idea as well. But that's a ways away. So we don't use BONNIE, in other words, for -- it's not as if a developer can say, well, I can't test so I'll just use BONNIE and that's what I'll -- that, no, that doesn't happen. We only use BONNIE in one of those two situations. Any other questions?

CO-CHAIR GEORGE: So if we are presented with BONNIE output at this point, you have reviewed it and determined --

MR. GOLDWATER: That's correct.

CO-CHAIR GEORGE: -- that it was appropriate?

MR. GOLDWATER: Right.

CO-CHAIR GEORGE: Okay.

MR. GOLDWATER: We've done the assessment to make sure the logic is calculated correctly and everything is structurally sound, yes. That is not a judgment that we would ask the

committee to make.

DR. BURSTIN: Just one potential comment and thanks, Jason, that was great. But part of what we're also trying to just make the distinction between is, the measures that are going to be used in some of the, for example, new measures to be used in CMS payment programs, you want to be sure they actually work in the EHR that people are using. We recognize that's a heavy lift at this point. So by putting things into eMeasure trial-use, as we talked about yesterday, there's an -- we're not saying it's endorsed, we're saying it's approved for trial use, please go out there, try it. Probably don't use it for those high stakes uses yet, because it's not ready.

And the ones that are kind of these legacy measures that have been around for a while, we're still trying to kind of go through that process and get them ready as best we can. But it is very much the idea that we recognize that even testing in an idealized simulated environment, certainly would have no recognition to anything I

practice with and probably would be pretty different just even where the things were structured or not may be very different.

MR. GOLDWATER: And to follow on that comment, it was actually something that was made by one of my colleagues yesterday, which is if the committee is debating over the testing or the formatting of an ECQM, then we haven't done something correctly. That when we do the assessment and give it to you, that's all been done. The only that this committee needs to focus on is the actual criteria for evaluating a measure.

MS. MARINELARENA: So before we get started today, like Jason said, we evaluate these eMeasures as two separate measures, but we're only going to evaluate the eMeasure part today. And then on the call, we'll evaluate the measures as the original paper measures. Because I don't know that the rest of the lead discussants actually evaluated the paper measure version of it, so we'll give you some additional time and we'll follow up with you on that. Not sure how clear it was in the

preliminary analysis that it was two separate measures. So technically we would have been voting on six measures rather than three. So we're going to do the eMeasure version today and then on the follow-up call we'll do the original version on the follow-up call.

CO-CHAIR KOTTKE: Okay.

MEMBER MITCHELL: I'm sorry, can I ask a question?

CO-CHAIR KOTTKE: Yes.

MEMBER MITCHELL: So the materials that we received for those of us who evaluated the eMeasures, was only about the eMeasure, correct?

MS. MARINELARENA: It's both in there. They do have -- right. So they provided the eMeasure version and the specs and all that in the attachments. But then there's also the tested information for either the registry. It was both in there, but it wasn't that clear and the preliminary analysis was sort of combined. So it wasn't that clear. So if you just did the eMeasure portion, that's fine. That's what we're going to

do today and then we'll follow up with you afterwards.

CO-CHAIR KOTTKE: Okay. The first measure of the morning is 0070, Coronary Artery Disease: Beta-Blocker Therapy - Prior Myocardial Infarction or Left Ventricular Systolic Dysfunction, from the AMA-PCPI. A brief update or description please?

DR. RADFORD: Good morning. My name's Dr. Martha Radford. I'm a cardiologist and I'm a member of the Executive Board of the AMA-PCPI. I've been doing quality performance measurement and improvement for 20 years. And I just want to say that these measures -- I'm going to talk about the three measures. This is the introduction for three measures. Coronary disease and heart failure are certainly very important, impactful diseases. These measures are adapted from measures that have been developed and in place really for 20 years, but endorsed by NQF since 2003. And these are the eMeasures and they're focused on out-patient care to a certain extent with some

in-patient input.

The evidence base on these measures is extremely strong and, again, been in place for at least 20 years. And any cardiologist would consider these motherhood and apple pie measures really. So, that's -- they've been tested extensively in regular records and as much as possible in e-records by the AMA-PCPI, which has done a lot of testing, or more than anybody else, in the e-environment. And I'll say -- I'll end my remarks there for the three measures.

CO-CHAIR KOTTKE: Thank you. Our discussants are Leslie and Judd. Leslie, are you the primary?

MEMBER CHO: Good morning, everybody. So this is the Coronary Artery Disease: Beta-Blocker Therapy - Prior Myocardial Infarction or LV Systolic Dysfunction with EF less than 40 percent. So before we get started, I just have a couple of questions for the Developers. The measure doesn't specify how long the beta-blockers should be used if you just had prior MI.

MS. TIERNEY: Yes. There is detail within the measure specifications and actually in the denominator that the prior MI has to have occurred within the past three years.

MEMBER CHO: I read that. But does that mean that if you had a heart attack three years ago and this is year four, that we're not going to give beta-blockers? They're not considered in the denominator?

MS. TIERNEY: It means you're not considered in the denominator, correct.

MEMBER CHO: Okay. The second question I have is that this measure has been endorsed since January of 2009, recent re-endorsement in 2012, and we still don't have the performance data from the eMeasure. Why is that?

MS. TIERNEY: So the eMeasure has -- Kim maybe you can speak to how long the eMeasure has been around.

MS. SMUK: So the eMeasure has been use, I believe there was a version available through the PQRS EHR option. And then it was evolved and it

took some different pathways, but the eMeasure is now available and in use in Meaningful Use. And because both of those are government programs, we don't necessarily have access to that data. And CMS doesn't publish the EHR data in the same fashion they do with the PQRS claims. So that is not widely available.

MEMBER CHO: So we are going to -- I just want to understand that we are going to talk about an eMeasure for which we do not -- we'll never have a performance gap information on the eMeasure? Or that will be very hard for us to get a performance gap from the eMeasure?

MS. TIERNEY: So I think Kim described it well. I think the Meaningful Use program currently doesn't provide performance data at all. So I'm not -- I'd say as the MACRA legislation rolls out and with the onset of the MIPS program, which is going to be an attempt to combine PQRS and Meaningful Use, there probably will be publically reporting of performance data on Meaningful Use EHR measures. But that probably won't happen until

2017 or 2018.

CMS occasionally will give us confidential performance data and so we have given you some performance data. They also publish experience reports as well and we've given you the data from that. But, again, they currently report that on the claims or registry versions of the measures and not necessarily the eMeasure versions of the measures.

MEMBER CHO: I don't know. I'm having a philosophical debate about endorsing a measure for which we don't have a performance gap. I understand the registry we will have a performance gap. But I'm a little ambivalent about not having a performance gap, I think you can all understand my angst here, right? So, okay. So my third question is, is why is active heart failure not included in your exclusion criteria? Beta-blocker use for patients who are in active heart failure regardless of whether they have had heart attack or not, EF less than 40, I looked through your exclusion criteria, it doesn't

specify active heart failure.

DR. RADFORD: I believe you're referring to decompensated heart failure. Again, this is a difficult concept in an e-environment. And this is meant to be --

MEMBER CHO: But not in the --

DR. RADFORD: -- a chronic care measure.

MEMBER CHO: -- ICD-9 or 10 world where you can actively code in an out-patient setting whether patient is in heart failure, currently in heart failure and cannot get beta-blockers.

MS. TIERNEY: So the other thing to add is that the PCPI methodology includes three broad types of exceptions. So there's a medical exception and we have hard-coded the examples that we've listed in the measure. And those are based on clinical evidence and the expertise of our work group who developed the measure. But there's also this other medical reason option and so certainly if a physician did not prescribe the beta-blocker because a patient was in decompensated heart failure, they could use the other medical reason

exception to account for that.

MEMBER CHO: My other question is, is that you specified three specific beta-blockers for your CAD. So in your report, in your performance gap that you listed for the registry, is that for the three specific beta-blockers?

MS. TIERNEY: So just to clarify, so for this measure, there's two separate patient populations. There's the patient population who had a prior MI within the last three years. And because the guidelines do not specify a particular type of beta-blocker and there appears to be a class effect among beta-blockers, we do not require it be those three beta-blockers for the prior MI population. For the other population with the LVEF less than 40, we do require that it be the three beta-blockers that are recommended in the guidelines. I think that recommendation only came out in late 2012 with the new guidelines from ACC and AHA for stable ischemic heart disease.

And so the data that we've reported on the performance is actually from the 2013

experience report, which would have pre-dated the guideline update and when we actually made the update to the measure. So the data that you have is more generic for just the prescription of beta-blockers in general and not specific to those three beta-blockers for that second population within the measure.

MEMBER CHO: The measure also requires that patients have two separate visits to the provider to have a provider-patient interaction before the beta-blocker use is considered. How is that time? What's the time specific between the two visits?

MS. SMUK: There's no requirement the two visits be separated by a given period of time. The only requirement we place on the two visits is that the two visits happened within the measurement period.

MEMBER CHO: The measurement period between 12 months of a heart attack?

MS. SMUK: No. Just the measurement period based on the program that it's implemented

in. So that's just in a 12 consecutive month period, were there two visits? Regardless of when the condition was the CAD or any other factors.

MEMBER CHO: All right. So let's move on to -- I'm think I'm okay with all the questions and the answers that the Developers provided. I want to thank you for that. So we can move on to the discussion about the Coronary Artery Disease: Beta-Blocker for patients who have prior heart attack or LVEF less than 40 percent. As the Developers have provided, there is strong evidence from the guidelines on patients with LVEF less than 40, numerous randomized control studies. And then also for patients with heart attack, the evidence is less clear after year one, but certainly there is some evidence up to year three. So I think on the Evidence aspect of the measure, I think it's very strong.

CO-CHAIR KOTTKE: Judd?

MEMBER HOLLANDER: I'm just going to ask, George sent around that article yesterday. I was on a train and not part of the discussion, so

I was going to actually ask George to comment on how he thought that played into this measure and let the Developers address that.

MEMBER PHILIPPIDES: Yes, thanks. So there is evidence and meta-analysis most of the last few years suggesting that beta-blocker therapy was shown to be much more effective in pre-reperfusion. So a big heart attack, LV dysfunction, non-revascularize I think is a big part of it. Now, when we change our practice and everyone gets revascularized, if you're left with normal LV function and open coronary arteries, the benefit is much harder to show, even earlier on.

Having said that, I think that the guidelines do point out studies from a mixture of people, reperfusion, ATF perfusion and their overall recommendation is, well, we don't know, go to three years. So I think to be consistent, I would probably just stay with the guidelines. My prediction is the next set of MI and unstable heart failure guidelines might tweak that further. But for now, I think we should go by the guidelines

as written.

CO-CHAIR KOTTKE: Okay. Thank you. Any other comment before we vote on Evidence? It sounds like the evidence is high. Let's vote on Evidence.

MS. IBRAGIMOVA: Importance to Measure and Report, 1A, Evidence Structure Process Intermediate Outcome, 1 High, only eligible if QQC submitted, 2 Moderate, 3 Low, 4 Insufficient. Can we revote?

MS. VICALE: Please just place your vote again. Thank you.

MS. IBRAGIMOVA: We have one more vote that we need.

MS. VICALE: Please ensure that you pointed it in the direction of Laura. We appreciate your patience, we're just having some trouble with the voting software. Bear with us, we're just going to work on getting this worked out.

MS. IBRAGIMOVA: Do we feel comfortable doing a hand vote? Okay. All of you voting High? Okay, High keep your hands up for me. One, two,

three, four, five, six, seven, eight, nine, ten, 11, 12, 13, 14, 15, 16. I got 16, 16. Okay.

CO-CHAIR KOTTKE: Thank you. Opportunity for Improvement. Leslie, please?

MEMBER CHO: So the performance gap was not provided for eMeasures. But performance gap based on the PINNACLE Registry still continues to show some performance gap between the years. It still is hitting around 70 percent. So there is still the room for improvement.

CO-CHAIR KOTTKE: Judd? Any further discussion on Performance Gap? Okay. Let's vote on Performance Gap -- Opportunity for Improvement, sorry.

MEMBER CHO: I hope when this measure comes back for re-endorsement, we will have some information about eMeasures. Because for us as a committee to approve a measure without understanding or having a performance gap, leaves, I think, both of us in not a good position.

DR. BURSTIN: Just so you know, Leslie, I've actually also contacted the folks at ONC and

see if we can directly get a feed. I mean, this is national data and I don't understand why there's confidentiality concerns about national level data. So we'll see if we can get that back to the committee.

CO-CHAIR KOTTKE: I need some instruction from NQF. We don't actually have evidence on the eMeasure. So is it Insufficient with exceptions, is that what we have to conclude? Okay. So we'll go ahead and do a hand vote.

MS. IBRAGIMOVA: Well, we got it up, so let's see if it works this time.

CO-CHAIR KOTTKE: Oh, it works?

MS. IBRAGIMOVA: Yes. Well, I don't know. We'll see. Importance to Measure and Report, 1B, Performance Gap, 1 High, 2 Moderate, 3 Low, 4 Insufficient. Well, we might have to revote, actually. We have one recusal.

CO-CHAIR KOTTKE: Don't we have two recusals?

MS. IBRAGIMOVA: Two recusals, yes. Can we just revote?

CO-CHAIR KOTTKE: Revote, please.

CO-CHAIR GEORGE: Yes, we mistakenly had an extra vote on that one. Please revote.

MS. IBRAGIMOVA: So the results are, 25 percent High, 75 percent Moderate, zero percent Low, zero percent Insufficient.

CO-CHAIR KOTTKE: Okay. Specifications, Reliability, and Reliability Testing. Leslie, please?

MEMBER CHO: So for Reliability, the eMeasure and registry specifications are not similar, is that correct? The Developers? They should be exactly the same?

MS. SMUK: It depends on the particular implementation and what that program requirements are. So while they capture the same clinical concepts and intent, a lot of it depends on the program and how they're implemented. Because the specifications will need to vary depending on the program.

MEMBER CHO: Depending on the EHR program, you mean? Like depending on whether it's

Epic or something else?

MS. SMUK: No, not necessarily by vendor. But based on program. So some of the -- like if you are implementing in a PQRS registry model, that would look different because they have different ways of capturing the clinical concepts rather than in an EHR. So that's where the variation comes from is based on, we call them PQRSisms. So every program has a different look and feel and way of capturing the same clinical concept just based on the program structure.

MEMBER CHO: I have a question about the exclusion criteria. Why is pacemaker an exclusion criteria for getting a beta-blocker?

MS. MARINELARENA: Now you're talking about the eMeasure version, correct? Because --

MEMBER CHO: Yes.

MS. MARINELARENA: -- that's the one that we're focusing on. Okay.

MS. SMUK: So in this measure, which is 0070, which is CMS 145, the exception is actually that if the patient has an AV block, the patient

cannot have a cardiac pacer.

MEMBER CHO: That makes no sense. That is -- this is why I keep on -- that's why I was asking you about the pacemaker thing. If they have an AV block, they should have a pacemaker for which then they can get a beta-blocker.

MS. SMUK: So the way that it's structured in our logic is that if you have an -- maybe Dr. Radford can help us with this.

DR. RADFORD: Well, again, if they have AV block, that's an active problem.

MEMBER CHO: No, that's not how it's said. If it's an AV block comma, I understand. Right? I'm an interventional cardiologist. I'm with you. But if they have a pacemaker, it's an exclusion criteria.

MS. SMUK: In the logic, it's saying, if you have AV block -- in order to qualify as an exception, if you have an AV block, you cannot have a pacemaker to qualify as an exception. If you have an AV block and a pacemaker, you no longer qualify as an exception.

CO-CHAIR KOTTKE: So, Leslie, it sounds like it's appropriate that if you have a pacemaker and AV block, you're no longer an exception and so you would be expected to have beta-blocker.

MEMBER CHO: Well, the exclusion criteria, if you go look at the worksheet. If you look at your worksheet, the Excel spreadsheet, it's listed as your exclusion. Yes. Look at your Excel flow sheet, please.

MS. SMUK: So you're looking at the value set spreadsheet?

MEMBER CHO: Correct.

MS. SMUK: Okay. So the value set spreadsheet cannot be used alone. These value sets are designed to be used in conjunction with the HTML/HQMF specification as Jason had alluded to. Those go hand-and-hand. And so just having a concept in the value set spreadsheet alone does not indicate then how it is incorporated into the calculation of the measure.

So in the HTML/HQMF piece, you'll see the logic and our logic says that, and AV block and

not cardiac pacer and not cardiac pacer. So we have it applied at two different QDM levels. And so our intent here is to say in order to qualify and meet exception criteria, you would have to have an AV block and not cardiac pacer as defined by the values in that value set spreadsheet. And then, and not cardiac pacer, as defined by other values in that value set spreadsheet that coordinate to that value set title.

MEMBER CHO: So it's just AV block and not pacemaker?

DR. RADFORD: Right.

MEMBER AL-KHATIB: So the wording then in this document needs to be clarified if that's what you're stating. I mean, I hear what you're saying that those are like and statements in the algorithm that you're using, in the EHR. But I don't think this comes across that way in the document that we have in front of us. The way we're reading this is that if a patient has a pacemaker, they get excluded. So that just needs clarification.

The other thing that I would add as an electrophysiologist is that there are different types of AV block. And so maybe being clear that you're talking about type two, second degree or third degree, again, with the ICD-10 codes, we will be able to get to that level of detail. Even in the ICD-9 codes, we have that level. But, for example, first degree AV block should not be an exclusion.

MEMBER CHO: Okay. So we'll move on to -- oh, go ahead.

CO-CHAIR KOTTKE: Liz? Turn on your mic.

MEMBER DELONG: You don't have to answer this question if everybody else understands, but I missed the previous discussion when you said -- the way I interpreted it is that this measure gets customized for every different program that uses it in terms of its definition and implementation?

MS. SMUK: I think the best way to explain this, so it's the same measure. It's designed to capture the same grouping of patients

and it's designed to be uniform in that fashion. But the way that the specification looks, feels, the formatting is what gets customized based on program. And how a medical reason exception is captured in one program may be different than how it's captured in another program based on either coding terminologies or how this is actually reported by a physician. So that's where the variation comes in is actually at the specifications level based on program requirements.

MEMBER CHO: My other question is, is that the reliability of this depends on how well somebody puts in their criteria. You know what I mean? Because you do some reliability testing, which you submit, it's variable between physicians who put in lots of information versus people who don't. So there is that huge sort of variation in the reliability depending on which data set, I guess, we're looking at, right?

MS. TIERNEY: So what I think you might be referring to and I think this is our, and I know

we're not quite at testing, but this is the testing data that we provided related to the reliability performed on a signal-to-noise analysis of registry data. And what I think you're referring to is the fact that the reliability varied based on the number of events. So, the average was relatively high, 80 or 60 or something like that. And at the average number of events, the reliability was high. But at the minimum number that we assessed, at 10 events, the reliability was moderate.

So I think that's what you're speaking about. Which is a little different than I think the question about the different implementations and how the specifications differ a little bit. And I think Kim described it well. The only thing I would add is that the EHR can allow for much richer data, so for that reason alone, the EHR specification includes a lot more details and specifics than you would see in the registry specification because the registry cannot accommodate all of the richness of the data and the

terminologies that are part of the electronic measure.

CO-CHAIR GEORGE: This just relates to the broad scope of exclusions that you have in I think all three of these measures. And I think we had a little bit of this discussion before, but it includes not just medical reasons, but patient reasons, including family situations. And the description of the exclusions says that it incorporates things that may not be relevant. And I know we've had some problems with this, particularly CSAC has had some problems with these broad categories of exclusions, for instance, income. And I'm wondering if you can speak to that because I do see it as being problematic.

MS. SMUK: So this is something that probably originated a while back, but I'll speak to the more recent way that this kind of evolved. And so when we were working through re-tooling measures, going from claims to eMeasures, and we were working with other developers as well, there was this -- the goal was to not have a tailored

medical reason, patient reason, and/or system reason, wherever they're applicable for every specific measure. Because then you run into an infinite number of value sets, et cetera.

But also the reason we wanted to pick these broad lists, like medical reasons, is, A, because we didn't want to tailor it for each measure, but also that it allows reuse across measures, across measure developers, but it also allows that ability of the physician judgment to be able to make the call. And so there are concepts that may not be relevant to a particular measure, but you would have to trust that the providers would not be reporting those.

And so, a lot of this has to do with the fact that you would have to customize that list for every measure and there's an infinite possibility of subsets. And you would essentially have value sets that would have one concept and your measure would get very lengthy and it would be difficult to implement for that purpose. So we've tried to have a list that could be widely applicable across

measures, across Developers, and across various clinical situations to allow for physician judgment, but then also from an implementation standpoint.

CO-CHAIR GEORGE: Can you give us examples of system reasons?

MS. SMUK: Let me pull up --

MS. TIERNEY: So system reasons are generally things that are outside the patient's control and the physician's control. So things that are -- we've counted financial reasons, like there might be concerns over, I don't know, a patient's ability to pay for a medication or something like that. So that would fall under the system reasons.

They are -- these three measures, but this measure in particular since this is the one we're talking about, we did do an analysis of the exception rates because of sort of the concern I think generally expressed their validity. And there was a study funded by AHRQ, which we've reported on in the testing section, where we looked

at the frequency of the exceptions and they were found to be valid. We also looked at the medical record to see whether they could be validated. And so we found that generally they were used infrequently and they were valid.

And I think for this measure in particular, the percentage of exception reporting was 6.2 percent on average. And I think in our testing report, we have information about how many of those were medical exceptions, 84 percent were for medical reasons, 12 percent were for patient reasons, and 2.9 percent were for system reasons. So the system reason is used relatively infrequently, but it is primarily for financial reasons, insurance, things like that. I mean, with other measures, there may be other reasons, like a fluke shortage or something like that, but for this measure that doesn't apply.

CO-CHAIR KOTTKE: Am I recalling correctly that you give the physicians feedback if they seem to be --

MS. TIERNEY: Yes.

CO-CHAIR KOTTKE: -- quite deviant in the number of exceptions?

MS. TIERNEY: Yes. So we do recommend that physicians receive reports on their exception rates and that they document the actual reason for exceptions in the medical record or in the electronic system. The implementation of that is -- because of our -- this is how we recommend our measure be used, but we are not a measure implementor, so the implementation of that is out of our control, but that is what we recommend.

CO-CHAIR KOTTKE: Okay. Judd?

MEMBER HOLLANDER: So it doesn't appear to me that you have a minimum reporting threshold and your reliability as you spoke to it is moderate, but is below the cutoff that NQF exists of 0.7, it's 0.65 if there's ten reporting events or fewer. So I have two questions related to that. Shouldn't this measure just exclude people that have had less than ten events since the reliability is bad?

And then the second thing is, you say two or more visits for a patient-provider

relationship. Most of these people have a primary provider, a cardiologist, and maybe another doctor. Who does the measure get attributed to if within a three period or a one year period, they've seen four doctors two or more times? And so I think from a primary provider perspective, the cardiologist drives the decision making on these drugs.

So I think it's important to have clarity as to who gets attributed for doing this. I don't know too many family physicians who are going to put someone on beta-blockers when the cardiologist says don't do that. Whether it's right or wrong. So I want to get clarity around the small numbers and who it gets attributed to.

MS. TIERNEY: So with regards to the small numbers, I think, again, this is somewhat -- I don't know if we could -- I certainly think we could explore whether or not there's anything we could include in our documentation that would recommend a minimum sort of number of events in order for the measure to apply. I have to admit

I'm not sure I've seen that with other measures, so I'm not sure how feasible that is. But I think that's something we could consider.

I will say, just in looking at, and I'm not trying to get into different criteria, but in looking at publically reporting, for example, I know in the recent proposal from CMS that they've said that they're looking at physicians publically reporting on measures where physicians have had a minimum of 20 cases. So that would -- it seems like maybe the field is moving towards a certain minimum number of cases in order to apply a measure to that person.

But that is somewhat out of our control. I think we could certainly take it back internally and to our advisory committee and see if there's anything we can do from an implementation standpoint related to the minimum number of events. But I'm not sure that I've seen that yet in sort of the measure community.

MEMBER HOLLANDER: Well, I think we've seen it in other measures. But the specific

relevancy to me right now is, it says 0.7 is the minimal accepted value and you give me a number that's 0.65. So if I don't know you can get above 0.7, I should say you don't meet reliability thresholds. And since I'm going to have to vote in two or three minutes, I would say that's an important point of clarification that we kind of need to know now because I can't vote on maybe you'll change it, maybe you won't.

MS. TIERNEY: Right. I mean, one thing I will say just that the average was 61. So I mean, I'm not -- I think we'd have to look back at our data to see how the range fell in terms of how many had sort of the minimum. But the average was certainly quite a bit higher than the ten. The other thing, I guess, and I'd look to NQF staff to maybe advise on this, what you're speaking to is the registry data that we provided. And I know that there was an emphasis to just look at the eMeasure right now, and so I'm just not sure how much the registry testing should come into this discussion. I just defer to you as a staff to

determine that.

MS. WILBON: Right. So just two clarifications. One, we don't generally have a threshold per se for reliability testing. I think 0.7 is something that is generally accepted I think in kind of statistical and methodological testing that tends to be an acceptable number. So I think we probably have that number somewhere as a guide. But ultimately, it's up to the committee to determine what threshold of reliability based on the testing data that's provided by the Developer, whether or not that's acceptable in consideration of the other information and data that's submitted.

In terms of the evaluation, we would like to try to stick to the eMeasure specifications at this point. I know that they're kind of mixed in there, the registry data and the eMeasure data may be mixed in there a little bit. So to the best of our ability if we could tease that apart and try to focus on the eMeasure, because we are going to have vote separately. So --

MEMBER CHO: Even though we're voting

separately, we're trying to pass the eMeasure without really any data from the eMeasure on Reliability. And so we have to actually use something, right? We have to use something. And so the something we're using is the registry unfortunately, or fortunately. So how -- so that is a quandary, right?

MS. WILBON: Yes.

MS. TIERNEY: So just to clarify, I mean, I think that this is what I think Jason described as a legacy measure. So, in terms of the requirements for reliability testing, my understanding is that we needed to provide testing from a live EHR, which we have. Ideally it would be from more than one system. Our testing data is only from one system, but when a legacy measure does not include information from more than one system, the BONNIE testing can help supplement that. So we do have testing on the eMeasure, it is presented in the testing attachment.

CO-CHAIR KOTTKE: Sana?

MEMBER AL-KHATIB: I'm trying to

actually think how that would play out within the EHR system. So at Duke we use Epic. So are you thinking that all these elements that you need for this Measure will come from like the problem list? From the bidding list? From -- where would they come from? And I'm mostly concerned about the EF, because there's not a data element for the EF that we can capture today. Can you help us with that? And this actually applies to all three Measures. So -- thank you.

MS. SMUK: Yes. And it's a good question. So a lot of it comes in with our specification does not dictate where data is stored, how it is stored, et cetera. So our HQMF specification is essentially a framework and a vendor implementor would have to either design their EHR system or dictate where in their system the data is pulled from. So, for example, some vendors, I mean, they may have to somehow develop a discrete field in their EHR to be able to pull an EF and to pull that to be able to extract that into -- to meet the criteria of the eMeasure. But

we don't dictate how a system is designed, et cetera.

What an HQMF specification does is just say these are all the data elements we need, here's the algorithm to be able to calculate it. And the HQMF is like the codes, those are the codes that need to be reported in order to meet a Measure or for appropriate reporting. But those aren't necessarily the codes that have to be stored in your system. A system could also implement local codes and then they would just have to map them to the appropriate codes that are used in the specification for reporting purposes.

MEMBER AL-KHATIB: Thank you.

MS. SMUK: You're welcome.

CO-CHAIR KOTTKE: Okay. We just --

MEMBER CHO: For the testing, you were talking about the EMR testing that you did on 134 patients? Is that what you're referring to, 134? Okay.

CO-CHAIR KOTTKE: So we got yellow carded here.

MEMBER CHO: Let's move on and let's vote on Reliability. So I think we've had a lot of discussion about reliability, I think at most we can only vote moderate, it is not high. Given the fact that -- all the issues we talked about today.

MS. IBRAGIMOVA: Scientific Acceptability of Measure Properties, 2A, Reliability, 1 High, 2 Moderate, 3 Low, 4 Insufficient.

CO-CHAIR KOTTKE: While people are voting, I jawboned our Epic people into putting the field in for ejection fraction so we can get it in our system.

MS. IBRAGIMOVA: So the results are zero percent High, 56 percent Moderate, 38 percent Low, six percent Insufficient. Which means it's in the grey zone?

CO-CHAIR KOTTKE: So that's grey zone, but we continue. Correct?

MEMBER CHO: So for --

MS. IBRAGIMOVA: It is in the grey zone.

CO-CHAIR KOTTKE: Validity?

MEMBER CHO: So for Validity -- are we moved on to Validity? So, it just basically indicates whether specification is aligned with, I think, the evidence and whether it's appropriately risk-adjusted with exclusions. We've talked a lot about exclusions and maybe changing the wording around it to make it better. So I think for Validity, I think it's moderate to high.

CO-CHAIR KOTTKE: Judd?

MEMBER HOLLANDER: Yes. I was going to fall on the other side of this one. And this one I have problems with validity that maybe you could help me through. So it did some BONNIE testing, but this was the question that I asked earlier on is that face validity it passes on the registry data. But on the eMeasure testing, the developers provided simple agreement for 134 patients, no kappa is provided, but it was only 82.8 percent simple agreement. That could be a kappa score that's horrible, I don't know what it is here.

If the numerator criteria, if you only

did EHR review, remembering that they only also used one EHR and two is recommended, if you do EHR review and manual review, you bump that number to 92.5 percent agreement, which tells me that you're picking up at least another ten percent in the manual process. And without knowing kappa values and true inter-rater reliability, I think those agreement numbers are quite poor.

I also think I have problems with the fact that the data's been out there and it's double-top secret, but when I, and I know that's not your fault, but when I look at these numbers it raises concern to me. And it's one EHR without good numbers. So I actually see this as problematic rather than acceptable unless you could tell me something else besides we also did BONNIE testing because I think this overrides the BONNIE testing.

MS. TIERNEY: So I'm going to ask our testing colleague, who's in the back of the room, if he can speak to the reliability information we have provided from that EHR testing.

CO-CHAIR KOTTKE: Can you come up to the mic, turn the mic on please?

PARTICIPANT: So the reliability, at least from our -- this mic? Yes, it's on. So the reliability from our standpoint, we found it to be at -- sorry. Sorry. The reliability testing that we performed, we found it at 90.3 percent. Okay, sorry. To address that, we found the reliability to be at 90.3 percent based on some of the testing that we did.

MS. TIERNEY: I mean, I think the other thing to point out is in determining sample size number, there are certain statistical considerations we take into account. So even though the number seems small, it is statistically significant. So I think that sometimes, especially when you look at the data that we've gotten from our signal-to-noise analysis, which is testing at the Measure score level, there's a large number of patients included in that and so the 134 can seem quite small. But the type of testing that's done is a different type of testing at the

data element level. So I think it's important to understand that distinction too and the fact that the 134 did produce statistically significant results.

MEMBER HOLLANDER: Well, so my issue isn't with the 134, it's with the 82 percent agreement in the 134. I mean, if we're going to use an eMeasure, then agreement should be high. And you can have 90 percent agreement and a kappa score that's poor. So at 82 percent agreement, it's my guess that if you gave me a kappa with a confidence interval, it would certainly overlap a range we wouldn't feel confident in using. And I don't have that data.

So I think, my gestalt without having the data that we specifically need is that it wouldn't meet the threshold there. And, again, it's not because of the 134, it's because of the likely kappa that would be around only 82 percent agreement. And we know there's at least an extra ten percent we pick up by manual review. We don't know how we could best find the other eight percent.

But it seems to me there's a lot missing here that's concerning.

CO-CHAIR KOTTKE: Sana?

MEMBER AL-KHATIB: I just want to make one comment here. I mean, I completely agree that we should be targeting higher agreement rates and kappa statistics, but I would remind the group that yesterday we approved Measures where the kappa statistic was 0.55, 0.6. Just so we know. I'd like us to be consistent.

CO-CHAIR KOTTKE: Any other discussion? Are we ready to --

DR. RADFORD: I would like to make a remark --

CO-CHAIR KOTTKE: Yes.

DR. RADFORD: -- as a private citizen here, not as representing the AMA-PCPI. So, my day job is Chief Quality Officer at NYU, so I do a lot of quality performance measurement. We trying to move e in all areas and we're participating in CMS's experimental EQR program, which is extremely enlightening. And just to point out that eMeasure

development is in its infancy. And we're not there yet. And NQF endorsement of the Measure, essentially the eMeasure concept, is an important step in that development.

There is no way that we can get fully reliable eMeasures without this iterative process that we're talking about. And it involves players from the provider community, the EHR vendor community, and the Measure Developer community in order to make this really work and really sing. But we are definitely not there yet.

MEMBER CHO: Should we not do a trial Measure? Should this not be a trial Measure until it works?

DR. RADFORD: Basically what I'm saying is, all EHR Measures that are endorsed right now are trial Measures.

DR. BURSTIN: Right. I think the distinction here is this is a Measure -- again, this is basically a legacy Measure of a program -- a Measure already in a program. So we can't call it trial-use. They have presented data that at least

gets it over that bar. But certainly I think everyone would concur, there's a whole lot more work to do to understand performance in EHRs and hopefully getting some data from ONC will help there.

CO-CHAIR KOTTKE: Liz?

MEMBER CHO: Helen, just maybe for all of our sake, when we're asked to endorse a eMeasure without data, like the high standard NQF data that we've all come to accept, then perhaps it should be a different criteria than what we currently have been using for the other Measures. Because I think all of us are trying to struggle with -- and we agree that eMeasure is where we're going, but until we get there, to approve it based on all things that we have, validity, it's difficult.

DR. BURSTIN: It's a challenge. I think it's something we can continue to talk about. I think at this point, with those Measures already in programs and already endorsed for other -- again, this isn't a new Measure. This is a Measure that's been around for a very long time. We have

a very good understanding of how it performs in other data systems. And the question is, how much is that leap from one data platform to another going to significantly change, I mean, not so much validity, just because I think validity in this instance, you do have face validity on the Measure itself, that at least gets it a moderate and reliability, you've got some evidence, at least from what they've provided, that it meets the bar, low as it may be, at least for those legacy Measures. But we'll take those considerations. Certainly I think it's a valid concern.

CO-CHAIR KOTTKE: Liz?

MEMBER DELONG: Well, I think there's an important distinction between the legacy Measure and this eMeasure that is in its infancy. And while it may perform very well as a legacy Measure where there is actual medical record review, we're talking about dependence on specifying fields in an electronic health record and that specification from one facility to another may be different and the data may be captured differently.

So you can't say that because it performs well at the paper level, that the same -- even though the concepts are the same, that it's going to perform well with this electronic algorithm. So it's not exactly saying, it works well here, so we should assume that it's going to work well there. It's not that simple.

DR. BURSTIN: Again, we agree. This is a really difficult space. I think we do have the eMeasure feasibility at least, demonstrating that those data elements can be found. They've done the assessment, that they're using accepted value sets. I mean, there are a series of at least check boxes along the way that give you a sense that it's getting closer. Would we love to have more data on its actual performance in multiple EHRs? Yes. And in some ways, keeping it out in the space helps us get there faster, I think was part of the argument we were hearing.

CO-CHAIR KOTTKE: I think --

MEMBER DELONG: I'd like to clarify what our --

CO-CHAIR KOTTKE: Can I just --

MEMBER DELONG: -- role is here. Because an endorsed Measure, I would think would signify something. And we're basically -- we're approving Measures on the basis of data that aren't quite up to speed but they may get there. And we're now being asked to approve Measures that have no real experience. I would think -- I don't understand what the hurry is. Why do we not wait for a Measure to actually demonstrate validity and reliability and be in use? I don't understand what endorsement means if it's not ready for show time.

DR. BURSTIN: I think you've raised two important issues. The first is the reason we decided to do trial-use. We recognize it is really, really hard to get multiple EHRs to test Measures. Many of them can't do it yet, they're not available yet. It is clearly the rate limiting step for getting some Measures to market that are eMeasures. We recognize that. That's the first thing.

The second thing is really about should

we call them endorsed? And this is really just the -- to be perfectly honest, this is an issue with these Measures kind of being pushed out very early on to meet early programmatic requirements. And the question is, how much do we expect of these early Measures that will likely change and get improved as EHRs change? And it is truly, I mean, I don't want to overuse the word, but it is a bit of a legacy.

Again, if you're not comfortable with that, we can further discuss it, but it is at least the policies we've agreed to for now across the work we do with the approval committee and others. And I think Jason may want to make a -- do you have your card up, Jason? You want to say something?

MR. GOLDWATER: Is this on? Great. So, to just follow with what Helen is saying and I'll make it very brief. The problems that you're bringing up are consistent problems in EHRs. This is not just germane to cardiovascular disease. That there are data quality issues in electronic health records and there have been for some time.

And those aren't going to go away at this point in time. That the reasons those exist are numerous, that there are unstructured fields in EHRs, that not all of the data is captured, that not all of the data is captured the same way, that different vocabularies are used depending upon what they're trying to capture.

I think it was an excellent point that the data richness in an EHR is vastly different than what is in a registry because a registry generally relies on claims. Claims are designed to do one thing, to pay out a provider based on the services that they're delivering. An EHR contains far richer, deeper vocabularies which give a better sense of not only the diagnosis and the procedure, but everything that went along with it.

So if we're going to evaluate an electronic Measure off the basis of data quality within the EHR, that is going to be very challenging to do because you're going to have some of these persistent problems until we get to a point of standardized data elements within the system

itself. Which has been, as Helen knows, a longstanding issue that we are still working on trying to rectify.

So, I think the issue in evaluating an eMeasure to determine its suitability for potential endorsement really comes down to is the data being captured appropriately? Is the data found within the system that can populate the Measure appropriately? Are the correct value sets there that represent the intent of the Measure as well as the intent of what is being captured? And is the logic calculating correctly so that the performance of the Measure is adequately being displayed?

Those things you can tell off the basis of the assessments that are being delivered. If we're going to focus heavily on data quality issues, it's going to be incredibly difficult to push any type of electronic Measure across.

CO-CHAIR KOTTKE: Right. If I could just comment, not as Chair. I mean, I wonder if we're really contributing to the problem of poor

data by endorsing a Measure that's not ready for endorsement. I mean, when I sat on the Preventative Services Task Force, we were asked to make decisions about like screening for prostate cancer without data simply because we were the experts. So are we contributing to the problem and saying, yes, it's good enough for the government? And then other people think that these are really good reliable Measures and in fact when we've sort of glossed over it. I don't know the answer.

MEMBER AL-KHATIB: Well, I mean, I think that I can assume that we all agree that we need eMeasures. I mean, there's no question that's where the future is. I think what we're struggling with is what criteria do we want to use to approve these eMeasures? And do we -- based on what we see, is that enough to approve the Measure? And you raised a good point regarding what will happen, let's say we all vote this Measure down today, what will actually happen to that Measure? How much will that set you back and is there a room for conditional approval or conditional endorsement?

MEMBER CHO: I'm still trying to figure out why we can't do a trial Measure? I understand this is an approved Measure, but we approved a Measure yesterday for trial. And I would like to trial this and see how things go. Because I firmly believe if this was a paper Measure, you guys would be out of here, we'd be done, you're thumbs up, we're out of here. But the problem we're having is this whole eMeasure concept. And so I would like to approve you, but conditionally on a trial basis. And I think for us to -- what is the hurry? I don't understand what the hurry is for the eMeasure.

MEMBER DELONG: And why even endorse it? I mean, as we discussed yesterday, we're fearful that we're not going to be here three years from now and we endorse it as a trial Measure and it comes back three years later and a new committee sees it as endorsed and rubber-stamps it.

MEMBER HOLLANDER: And I think Tom's point's really important. Is that at some point, this is -- we're left with the process as the

process is right now. And I actually, frankly, there are slides with very specific wording that I'm voting on and I personally can't vote to put this through for the reasons that are discussed. And I feel uncomfortable, and we've done this in the past in other rounds, where we vote for something that doesn't meet the letter of what we're voting for.

And I think we do more to fix the system if we vote it down and either NQF has to change what they're asking us to vote, call it a trial Measure or maybe it gives NQF power to go to the people that do control these things in the federal government and say EHRs aren't ready for prime time because there's no inter-operability and there's no data element coding that needs to be the way we need it and force change. But if we accept mediocrity, what we get is mediocrity.

And I just -- the other thing I'm going to add is, we do have some data here. This is not a total void of data, the data's actually not good. So this is a little bit of difference than when we

just have no data. Here, again, 82 percent agreement is a kappa value that may be well below 0.4, it may be nowhere near 0.55, and we don't even know those numbers. So without those numbers and with a statistical guess that really doesn't even meet the lower limits of the bar, there's multiple issues.

CO-CHAIR KOTTKE: Judd is left, speak to us.

MR. GOLDWATER: Okay. So I'm going to make this easier by just sitting up here rather than having Marcia run back and forth with the mic. A couple of issues and then -- and I think we fully understand what you're saying. And I understand that the data quality issues that are systemic in EHRs make this awfully challenging. And perhaps then it does ask us then to perhaps reexamine the criteria of evaluating eMeasures at this stage. Just a couple of points.

Number one, you're not going to resolve data quality issues in EHRs in this committee. That is not going to happen. Those problems have

been around for two decades and they're not going to go away tomorrow and they're not going to go away by any decision that you make. Those problems, because of the lack of standardization of data and the lack of inter-operability between systems, are pervasive, they are systemic, they are long-lasting, they are not going to end.

There's nothing you're going to do or any committee is going to do and there's certainly nothing we can do that is going to influence vendors and organizations to suddenly become inter-operable. I have been doing this for 22 years. I have been having the same arguments over and over and over again about how we get to a point of inter-operability and no one is listening to me. Which is fine, I'm used to that, nobody listens to me.

The second issue is, while I understand that there's perhaps a desire to want to put this into the trial-use program, but let me explain I think why that becomes somewhat problematic. And it's not because the intent is not reasonable, it's

because of how it's going to be perceived outside of this group. Which is, you're talking about a Measure that's already endorsed and a Measure that's already being used and a Measure that has been used effectively for a significant period of time.

Trial-use was not designed to put Measures in that are already being used, that are already being endorsed. I said this yesterday. It's not as if we don't have enough data, we can't test the Measure adequately, so let's put it into trial-use. The reason being is that, that will generate an incredible amount of resistance outside of this group. Because people will say, why are you putting a Measure for trial use that's already being used, that's already endorsed, that already has an NQF number, that's showing to be effective? And if we say that the issue is because we have concerns over the data quality and how that Measure is going to be performed, the question that's going to follow that is, was there testing that was done on the Measure? And there was

testing done.

Now, if you don't think the testing was adequate, then, yes, vote accordingly. Nobody's going to -- I'm not going to tell you how to vote. That's not my job or anybody else at NQF is to tell you how to vote. If you don't think the data is adequate to endorse the Measure, then vote accordingly. But we can't say -- it's going to be very difficult on all of us, especially on NQF, if we say we don't have enough data, we are concerned about the quality of the data, we are uncomfortable with moving past validity because of the results that we're seeing, so let's put it into trial-use.

Because I would image that Patrick Conway and Kate Goodrich, who are the ones that are pushing for this, and others will be like, why are you taking an endorsed Measure that's used in national programs and that has been around for two decades and you're putting it into trial use because we're just respecifying this to be electronic?

CO-CHAIR KOTTKE: So --

MR. GOLDWATER: Why are you not just simply rejecting the Measure or --

CO-CHAIR KOTTKE: But this Measure is not endorsed. This eMeasure is not endorsed by us, by NQF. Is that not correct?

MS. SMUK: The prior version of --

MR. GOLDWATER: The prior version, correct.

MS. SMUK: -- the Measure is endorsed.

MR. GOLDWATER: Right. Yes.

MS. SMUK: Correct.

MR. GOLDWATER: The prior version. This one is not, no.

CO-CHAIR KOTTKE: The prior eMeasure was --

MS. SMUK: No.

MR. GOLDWATER: No. The prior --

CO-CHAIR KOTTKE: -- there's a prior eMeasure? I mean, there's two different Measures, right? There --

MR. GOLDWATER: That's correct.

CO-CHAIR KOTTKE: -- is a register

Measure and an eMeasure.

MR. GOLDWATER: That's correct.

CO-CHAIR KOTTKE: The eMeasure is not endorsed --

MR. GOLDWATER: But the --

CO-CHAIR KOTTKE: -- it's a new Measure.

MR. GOLDWATER: -- initial Measure, the 0070, which is what I think -- that was endorsed, was it not?

CO-CHAIR KOTTKE: Yes.

MS. SMUK: Right. And --

CO-CHAIR KOTTKE: Yes.

MS. SMUK: But back then --

MR. GOLDWATER: But that --

MS. SMUK: -- there wasn't the distinction between -- when you got endorsed, it was only for one data source and I believe at the time, there also wasn't requirements on HQMF. And we did submit a PCPI e-specific which adhered to QDM, et cetera, it just wasn't in the HQMF format. So we did have an early eMeasure specification before HQMF was a requirement and before there was

a distinction of, you only get endorsed for a particular data source. So that is -- that prior endorsement was before any of these standards came about.

And one other just note is that these three cardio Measures are in Meaningful Use. They do have a lot of eyes on them. There are public platforms where issues either with the specification, issues with the clinical content, it's called JIRA, and anybody who has an issue with the Measure can go there and either ask for inquiries or ask for clarifications on the specification, on the intent of the Measure, I mean, it's kind of open for fair game, or just simple questions on the standards that surround them.

And our team was talking earlier this week and we're like, these Measures are actually the ones that we get the least number of questions on. Implementation wise, they're the most straight-forward based on their data elements and clinical concepts. These are the least in the grey

area. So that's something to put out there is that implementation wise, we don't get nearly as many questions on these Measures that we get on some of our other Measures.

DR. BURSTIN: So just in terms of the path forward, because obviously we're going to go through this again for the next two Measures. Well, I mean, the same issues will emerge. We won't? Okay. So, for better or worse, our current policy assumes that for these legacy Measures, this is sufficient. We would ask you to vote on what our policy says. But we've clearly heard your concerns here and we'll take it back and see -- we'll also reflect this very clearly in the report. Again, this isn't the end of the game, you guys know this, you've been around this block. This goes out for comment, we can extensively include in the report the significant concerns raised about this. And if you vote it Insufficient, we'll have to justify that.

But, again, keep in mind according to the letter of what we currently allow, this does

in fact meet that bar. We all recognize we'd like it to be a higher bar and we're hoping that over time Measures will move in that way. And we'll certainly think more about whether there are opportunities to call other Measures trial Measures.

CO-CHAIR KOTTKE: So are we ready to vote? Okay. Voting on Validity.

MS. IBRAGIMOVA: Scientific Acceptability of Measure Properties, 2B, Validity, 1 High, 2 Moderate, 3 Low, 4 Insufficient. Still need one more vote. Oh, thank you. So the results are zero percent High, 31 percent Moderate, 44 percent Low, 25 percent Insufficient. So it does not pass.

DR. BURSTIN: I would recommend that you just finish the evaluation of the Measure and not stop it here just because I think there's so many issues at play we'll need to follow up on.

CO-CHAIR KOTTKE: Okay. Feasibility?

MEMBER CHO: I think we can just vote, no?

(Laughter.)

CO-CHAIR KOTTKE: Judd?

MS. IBRAGIMOVA: Okay. Feasibility, 1 High, 2 Moderate, 3 Low, 4 Insufficient.

MS. MARINELARENA: Before we vote, if we could just have a little bit of a discussion or some -- before we -- for the record and for --

MEMBER CHO: Okay. For the record, I think we talked a lot about the electronic source being not reliable at times because of the, A, the richness of the data, B, the bad data that's sometimes going into electronic medical records. And I think that, that speaks greatly to how this Measure will be -- the performance gap numbers in the end.

DR. BURSTIN: Right. But that's not feasibility. Feasibility is really just are the data available in electronic data sources. So I think we've had that discussion under Scientific Acceptability. This is really about is the data source feasible.

MEMBER CHO: Right. And I'm --

MEMBER AL-KHATIB: I just want to make one comment, one real quick comment. I mean, I really would like to remind everybody most of the performance Measures that we have in use now use claims data. And so are you telling me that claims data are better than EHR data? I'm not sure.

MS. WILBON: I also just want to point out and maybe Jason can clarify, the use of the BONNIE tool that you guys also used the BONNIE tool is one tool that is used to try to help with feasibility and just to make sure that the data is feasible to capture from an EHR in terms of identifying the right data. So I just want to bring that to your attention that, that is one purpose for using the BONNIE tool is to help demonstrate that there is some level of feasibility in the electronic capture of the data. So did I -- I just want to make sure, Jason, that I characterized that correctly.

CO-CHAIR KOTTKE: Yes, Judd?

MEMBER HOLLANDER: So the questions we're asked to answer for the committee under

feasibility, which I think are worth reading because this is what we're voting on, are the required data elements routinely generated and used during care delivery? Are the required data elements available in electronic form? Is the data collection strategy ready to be put into operational use? And if an eMeasure, does the eMeasure feasibility scorecard demonstrate acceptable feasibility in multiple EHR systems and sites? And so those are the things we're voting on.

CO-CHAIR GEORGE: And did we hear what that scorecard showed?

MEMBER CHO: The BONNIE scorecard actually was very good. It was like 100 percent.

CO-CHAIR KOTTKE: And so that combination of Epic plus BONNIE is the two systems?

MEMBER HOLLANDER: But it does say multiple EHR systems in the questions and BONNIE's the substitute for it because they didn't --

CO-CHAIR KOTTKE: Right.

MEMBER HOLLANDER: -- cover that.

CO-CHAIR KOTTKE: By criteria, it meets criteria with multiple. Okay. Yes, Linda?

MEMBER BRIGGS: I have a little bit different take on the BONNIE results. While they showed 100 percent agreement, only 82 percent of the data element concepts were there. So if that's true, then there's some missing data that they're not capturing. Yes, we got 100 percent agreement in the way it was tested, but not all the elements were available. That's my take on it. Anybody else?

MS. TIERNEY: So can I address that question? So, I think what you're referring to is the coverage. And BONNIE is a test of the Measure logic and so we test all the different logic pathways within the Measure. And so sometimes when you don't see 100 percent coverage, it's because we might not test every single pathway, for example, for the exceptions. We have a number of examples, we might just test one pathway, but not every single one we've identified in the Measure because we've still shown that the logic works when

you use exceptions, we just haven't tested every single variable that's within the Measure specifications. I think a colleague of mine on the phone has a comment too about feasibility. Meredith?

CO-CHAIR KOTTKE: Yes. Go ahead Meredith.

MS. JONES: Thank you, Sam. Can everyone hear me?

MS. TIERNEY: Yes.

CO-CHAIR KOTTKE: Yes.

MS. JONES: Okay. Thank you so much for the opportunity to comment. My name is Meredith and I'm with the PCPI. Sam's explanation of the BONNIE tool is correct. Just a quick clarification. It's not an EHR itself. It's a separate tool that will hopefully interact with the EHR. And as Sam said, the 82 percent coverage rate is really testing the logic pathways of the Measure. And we are working to get that up to 100 percent, we just haven't tested all of the pathways within the Measure exceptions.

So, for example, we have a number of diagnoses active that could be an appropriate exception, but for the sake of testing the logic, we've only tested one. Because it is simulated data. We are working to bring this up to 100 percent. It's just adding more pathways within the tool itself. With that, we have found that all of the simulated patients we've put into BONNIE are in agreement. So it is in 100 percent agreement, meaning we have tested the logic sufficiently.

Another comment on feasibility that I just want to bring up, is we did include a feasibility scorecard with our Measure submission. And within that, we have shown that the data is available, accurate, and is meaningful and thoughtful to the physician's workflow. So we have spoke with physicians and cardiologists about these data elements, including the numerator, denominator, and exceptions variables, and they have indicated to us via their EHR that these variables, data elements are readily available and do not cause undue burden to capture. And if we

could just look at the scorecard, we provided threes across the board. Thank you.

CO-CHAIR KOTTKE: Thank you. So ready to vote on Feasibility?

MS. IBRAGIMOVA: We're just missing two -- one more vote. Oh, we've got it. So the results are zero percent High, 63 percent Moderate, 13 percent Low, 25 percent Insufficient. This is not in the grey zone.

CO-CHAIR KOTTKE: I'm sorry, what did you say about grey --

MS. IBRAGIMOVA: It is not in the grey zone.

CO-CHAIR KOTTKE: Yes. Okay. Usability and Use?

MEMBER CHO: So currently, it is being used by PQRS in the incentive program, EHR incentive program, and in the quality improvement with benchmarking, the PINNACLE Registry. So I think for Usability, it's currently being in use, so I think it's high.

CO-CHAIR KOTTKE: Judd?

MEMBER HOLLANDER: So I have a question because I'm confused. Because is that usability the eMeasure or is that usability the registry? I think it's -- it is the eMeasure? Okay. And then, talk to me about because I don't understand this, it says NQF's Measure Application Partnership reviewed the Measure and has the following recommendations, and none of them actually support use of the Measure. What does that mean? Why is that here? And how should we interpret that?

MS. WILBON: So that was just an FYI. So the MAP -- just in terms of how the Measure is being used and being considered for use in other applications, so our Measure Application Partnership, as Melissa described yesterday, they make recommendations to HHS on which Measure should be used in particular federal programs. So this Measure was considered in that context, but for use in a specific program. It wasn't about -- it's not about endorsement. It's not about use in any particular setting specifically, but just about a particular federal program. So that is just

context in terms of how the Measure's being considered for use outside of our sphere. Hopefully that helped clarify.

CO-CHAIR KOTTKE: So it sounds like the Measure is both being used and is usable? Vote?

MS. IBRAGIMOVA: Usability and Use, 1 High, 2 Moderate, 3 Low, 4 Insufficient Information. So the results are, 44 percent High, 50 percent Moderate, six percent Low, zero percent Insufficient Information. It passes Usability and Use.

CO-CHAIR KOTTKE: So I need a little instruction here on the overall. Do we vote on overall since it didn't pass Validity?

MS. WILBON: So I think we should -- since it did -- technically we generally would stop evaluating the Measure after Scientific Acceptability because it did not pass that criteria. So generally sometimes because there was a lot of process left and sometimes the Measure continues to be discussed, we do like to continue with the other criteria so we don't have to go back

and rehash the discussion in case the conversation continues with public comment. But let's hold off on a final recommendation because technically with the vote on the Scientific Acceptability, the Measure did not pass. So we'll just move on to the next Measure.

CO-CHAIR KOTTKE: So we're only two Measures behind.

MS. VICALE: Thanks, Tom. I did want to make a note to everyone, we do appreciate the robust conversation and we understand it's been rather arduous. However, even though we had originally scheduled a 10:15 break, we would like to continue on with at least the next Measure and keep going since it's a robust conversation. So we do appreciate your patience with that. Thank you.

CO-CHAIR GEORGE: And Developers want to make any comments about this Measure? I know you sort of addressed them all three to begin with. Anything other --

MS. TIERNEY: I don't think so, I think Dr. Radford gave a good kind of overall overview

of the Measures.

CO-CHAIR GEORGE: Okay. We'll go on to Joel and Mladen.

MEMBER MARRS: All right. So this Measure is Heart Failure and the use of Angiotensin-Converting Enzyme Inhibitors or ARBs for LV Dysfunction. And so similar to the first one we discussed, we're just focusing on the ACE/ARB therapy. And so to start off with Evidence, just like it was mentioned with the first one we discussed this morning, there's a tremendous amount of evidence and guideline recommendations for this Measure.

CO-CHAIR GEORGE: Any other comments?

MEMBER VIDOVICH: Very heavily researched field.

CO-CHAIR GEORGE: All right. We'll vote on the Evidence.

MS. IBRAGIMOVA: Importance to Measure and Report, 1A, Evidence Structure Process Intermediate Outcome, 1 High only eligible if QQC submitted, 2 Moderate, 3 Low, 4 Insufficient. So

we're missing one in the room. So the results are 88 percent High, 13 percent Moderate, zero percent Low, zero percent Insufficient.

CO-CHAIR GEORGE: We'll move on to Opportunities for Improvement.

MEMBER MARRS: So Opportunities for Improvement in Performance Gap, the Developers submitted information from PQRS data from 2010 to 2013. They reported just under an 80 percent rate of meeting this criteria in that system. And so still designating that there is a performance gap. They did go into talk about disparity issues and a need to further evaluate some of those pieces across multiple disparities that aren't necessarily publically reported at this time.

CO-CHAIR GEORGE: Any comments on the Opportunities for Improvement? All right, we'll vote on the Opportunities.

MS. IBRAGIMOVA: Importance to Measure and Report, 1B, Performance Gap, 1 High, 2 Moderate, 3 Low, 4 Insufficient. The results are 25 percent High, 69 percent Moderate, six percent

Low, zero percent Insufficient.

CO-CHAIR GEORGE: All right. We'll move on to the Reliability and Specifications.

MEMBER MARRS: So related to Reliability, they reported their reliability testing score that came out to be 0.94, with a sample of, I think it was 1,300 patients that they tested. Or 1,244 samples that they actually tested and showed a reliability score of 0.94. And then when they actually used the cut point like they did similar from the first Measure with that minimum of ten, reliability dropped to 0.83, but still designated high reliability overall for the Measure.

One of the questions that came up is how they were designating the denominator? The denominator says current or history of an EF less than 40 percent. And so the numerator specifies an ACE/ARB therapy in a 12 month period, but there was a question of what's the period range for the denominator? Is that within the same 12 month period? Or is that any historical EF less than 40

percent?

MS. TIERNEY: So it is any historical EF less than 40 percent. And our work group discussed, I think, a comment that I saw in the notes about should maybe the Measure focus on just a current EF less than 40 percent? And the feeling was that because these agents can sometimes normalize EF, that focus should really be on any current or prior EF less than 40.

Additionally, I think there is another Measure, I think it's going to be presented later today, about a LVEF assessment in patients with heart failure. And I know the guidelines recommend assessing it with a two-dimensional echo on initial evaluation, but not necessarily recommending the routine assessment. So I think the Measure's also consistent with that by focusing on the current or prior.

MEMBER VIDOVICH: There's just one thing I would just like to point out. This is the similar quandary we had with beta-blockers. We know that acutely and in short-term ACE inhibitors work, we

don't know three years, five years, ten years later. I mean, most of us clinically do continue them based on the assumption that and anecdotal evidence that if you take them off, the EF reverts, but we don't have data on that. I'm not aware that there's availability.

CO-CHAIR GEORGE: Sana?

MEMBER AL-KHATIB: Yes. I just have a quick question about one of the exclusion criteria, the marked azotemia. And I have to admit that I haven't heard that term, azotemia, in so many years. But how are we defining that? Is there a creatinine value that -- because in the EHR, nobody's going to say the problem less marked or something like that. Maybe with ICD-10, that will help. But can you help us understand how you're defining that?

MS. TIERNEY: Sure. So I'll have Kim explain how we're defining it or maybe not defining it as the case may be. But I will say that where it came from was when these Measures were last developed with a work group in 2009, and we do

complete annual updates based on new guidelines that are released, but when they were completed in 2009, we decided to include as examples of exceptions, medical reason exceptions, things that were specifically mentioned in the guideline. So that azotemia was particularly mentioned in the 2009 heart failure guidelines. I didn't look at the new ones to see if it's still in there. But that's where that came from, even though it may be an outdated term. But Kim can speak to how we might capture that.

MS. SMUK: So in some of the earlier specifications, there were codes that were used to capture that. And we found over the years and through feedback that having it hard-coded was not necessarily a valuable piece and that something like that, if it was going to be reported, would actually be better reported through a medical reason using a contraindication, et cetera. So it's not captured as a discreet stand-alone data element in our eMeasure specification. And rather we would guide people to report it through using

a medical reason.

And so this is just part of the evolution of an eMeasure specification. And when we saw this question in the materials that were distributed and we actually went back and looked and it was in a much earlier version of a specification, but through the annual updates that we do to our specifications, we did decide to consolidate some of the actual examples. But there's still the ability for a physician judgment through the medical reason value set.

CO-CHAIR GEORGE: Any other discussion on Reliability? All right, we'll vote.

MS. IBRAGIMOVA: So Scientific Acceptability of Measure Properties, 2A, Reliability, 1 High, 2 Moderate, 3 Low, 4 Insufficient.

MS. VICALE: Are we missing one more vote?

MS. IBRAGIMOVA: So the results are, 18 percent High, 82 percent Moderate, zero percent Low, zero percent Insufficient.

CO-CHAIR GEORGE: We'll move on to Validity.

MEMBER MARRS: So from a Validity standpoint, there's definitely alignment with the evidence recommendation for the use of ACEs and ARBs in this population and the Measure itself. The Developers submitted that it met face validity requirements and then they actually did the same sample test of that 154 patients that we talked about on the previous. And there was 96 percent agreement from using the manual evaluation versus EHR assessment.

CO-CHAIR GEORGE: Any comments on the Validity? If not, we'll vote on Validity.

MS. IBRAGIMOVA: Scientific Acceptability of Measure Properties, 2B, Validity, 1 High, 2 Moderate, 3 Low, 4 Insufficient.

MS. VICALE: We need one more vote, excluding Liz. Has everyone else voted?

MS. IBRAGIMOVA: Yes. Let's just try one more time.

MS. VICALE: Okay. We're going to do it

again.

MS. IBRAGIMOVA: So the results are six percent High, 88 percent Moderate, six percent Low, zero percent Insufficient.

CO-CHAIR KOTTKE: So, pardon my naivete, but has everybody just run out of powder? Or are they exhausted? Or is there something different between Measure 0070 and 0081?

MEMBER HOLLANDER: To me, the big difference is the data quality here. The data was 80 percent agreement as compared to 90-something percent agreement. I ended up personally giving this a Moderate rather than a Low, even though I still don't have kappa and I don't know what percent agreement actually really means. But at least it was a number that makes some sense. Eighty-two percent agreement when we know we do better in Measure 0070 with another ten percent by manually reviewing it doesn't hit the evidence bar to me. So I saw them as entirely different things.

CO-CHAIR KOTTKE: Anybody else care to offer a hypothetical explanation without revealing

your thoughts?

MEMBER MARRS: I guess I didn't, in the interest of time as well, discussing the first Measure extensively, I didn't feel the need to discuss some of the limitations and some of the other issues that we had already discussed and so just kind of highlighted some of the key pieces from a Reliability and Validity standpoint. But I think those same limitations from an EHR still exist with this Measure.

CO-CHAIR KOTTKE: Gerry?

MEMBER MARTIN: I voted the same on both times thinking that it was the same issue. Even though the data seemed different.

MS. MARINELARENA: So just to clarify on the last Measure, it failed on Validity? Yes. Okay. Just so we're clear.

CO-CHAIR GEORGE: Does anyone feel a need to revote on Validity? On the one that we're talking about? Okay. Then we'll move on to Feasibility.

MEMBER MARRS: So from a Feasibility

standpoint, all the pieces that would funnel into this are captured in an EHR and so easily collectible. And so overall, medical criteria of feasibility and actually all the Measure elements would be available. And so no major issues from a Feasibility standpoint that I can see.

CO-CHAIR GEORGE: Any comments on Feasibility? If not, we'll vote on Feasibility.

MS. IBRAGIMOVA: Feasibility, 1 High, 2 Moderate, 3 Low, 4 Insufficient. So the results are 22 percent High, 78 percent Moderate, zero percent Low, zero percent Insufficient.

CO-CHAIR GEORGE: We'll move on to Usability.

MEMBER MARRS: So from a Usability standpoint, the Measure is currently being used from a PQRS standpoint, a Meaningful Use standpoint, and the PINNACLE Registry successfully. And so no major concerns from a Usability standpoint.

CO-CHAIR GEORGE: Any discussion on Usability? We'll vote on Usability.

MS. IBRAGIMOVA: Usability and Use, 1 High, 2 Moderate, 3 Low, 4 Insufficient Information. So the results are 50 percent High, 44 percent Moderate, six percent Low, zero percent Insufficient Information.

CO-CHAIR GEORGE: Any other final comments? Liz?

MEMBER DELONG: I think Tom implied we're being a little inconsistent and I'm concerned that we're a little inconsistent. I mean, I think the discussion we had on 0070 was along the lines of this is not ready for prime time. The eMeasures as, is your name Jason?

MR. GOLDWATER: Jason.

MEMBER DELONG: Jason pointed out, electronic health records are not standardized across sites or across data systems and they have not yet been proven. So although there have been information supplied that lead us to think that these Measures can be valid and reliable and whatever, they're still fraught with the problems that were mentioned earlier. So I wonder how we

can -- I just don't think we're being consistent if we're saying all those problems exist for the first Measure, but they don't exist for the second Measure.

MEMBER CHO: I will say this for 0070, which is that it is a little bit more complicated than this Measure because it has many reiterations. So it has, you have to have an MI, you have to have had an LVEF less than 40, it's actually -- this is a much simpler Measure I think, because it's a heart failure, EF less than 40 percent. And so I think it's not the same in my mind.

MEMBER DELONG: EF less than 40 percent is incredibly difficult.

MEMBER CHO: I think you were all with us last year when we voted to have EF measurements as part of our -- didn't we vote for a Measure that mandated that heart failure patients have an EF on the chart last year as one of our Measures? Yes. So hopefully it will be better.

CO-CHAIR KOTTKE: So maybe what this is, is an example of the clinical strategy not to do

exactly the same thing all the time so at least you're not always wrong.

DR. BURSTIN: Just to weigh in briefly, it sounds like there was actually some discussion of the actual testing results on the one EHR at a higher degree of comfort. So I guess the one question might be, potentially offering just to come back and better explain or offer to show data on the first one since we haven't wrapped it up yet and see if we can just be consistent going forward. I think there was some -- just see if there's some additional way to make that work.

CO-CHAIR GEORGE: Judd?

MEMBER HOLLANDER: I think to Liz's point and Tom's inferences, I guess we'll now call it, is I don't think we get to vote on whether eMeasures as a whole are a good thing or a bad thing at this point in time. But it might actually be worth having in the report if the sentiments around this table are that eMeasures at least in the cardiovascular world are not ready for prime time and if we as a group think that they shouldn't be

moving forward until there's data and we shouldn't convert Measures that have been successful in another form to eMeasures until we can prove the eMeasures work as well or nearly as well as the other form. Like I think those are the sentiments I feel in the room.

We don't get to determine NQF policy, and we don't formally get to vote on that, but if we feel that way, it probably is worth having that documented in further discussions. But I think my votes were different between the two measures because I saw data to be different within the body of them. And I think that's what I was asked to vote on, so that's how my votes were different. I don't know whether Helen or anybody else thinks it's worth us having some kind of formal poll about whether we think eMeasures are ready for prime time or not.

DR. BURSTIN: I suspect Sana may say the same thing, but I'm not sure I heard that as the general sentiment. What I heard was specific concerns raised about the requirements for legacy

eMeasures that was reflected in the concerns about that measure. And we'll raise those concerns. I'm not sure we're ready to put something more global in or get into that discussion. But concerns heard.

MEMBER AL-KHATIB: That's exactly what I wanted to say. I don't know that, that represents the sentiments of everybody in this group. I certainly want to see some eMeasures endorsed and out there. And it would have been perfect, and maybe that's feedback for the NQF group, to maybe have something where either conditional endorsement or some trial track or whatever that might be, so that we all feel comfortable that we have enough data to support them. But I certainly don't want to put a hurdle in the road of development of eMeasures. We definitely need them.

CO-CHAIR GEORGE: Leslie?

MEMBER CHO: I think it speaks to the fact that if an eMeasure is simple, like this one is, and not so complicated as 0070 was, in terms

of the and, and, the multiple different things, I think eMeasures can work. For a simple measure, I think eMeasures can work. So I don't want to make a broad statement about eMeasures not being good.

MS. SMUK: So one thing that has come up a lot in our community is that when people look at a measure, they determine its complexity based on the lines of logic. And what -- a lot of the responses that have been given to us is that you think that because you're a human, but when these things are done electronically, they're not complicated at all. Because it is a computer system that's computing these and calculating these.

And so the lines of logic shouldn't necessarily take into account the complexity, et cetera. Because when you look at these three cardio measures side-by-side, they share a lot of the same data elements. One may have one additional data element, and one additional data element shouldn't add an overdue amount of complexity. Because they do share a lot of similar

--

MEMBER CHO: Then maybe you can explain why the difference in testing? So maybe that's something for you to come back to later.

MS. SMUK: Yes. Perfect.

CO-CHAIR GEORGE: Tom?

CO-CHAIR KOTTKE: Yes. I just -- I think I sort of share Sana's -- I mean, I think these should move forward. And when I sat on the IOM committee on cardiovascular surveillance, really tried to get a paragraph in there saying that with the EHR and with the Affordable Care Act, the EHR ought to be a real-time census and should really be a way that we can cheaply monitor.

And so I'm very concerned that there's something back behind that we're neither supposed to talk about or bring up here that's appropriate to bring up here, that we're inhibiting the development of further information systems. And I think -- I would like to see these things move forward and I think eMeasures are quite valuable and work should continue without us inhibiting

them. But on the other hand, there are still some issues.

CO-CHAIR GEORGE: Any other comments before we vote on endorsement of this eMeasure? All right.

MEMBER JAMES: Yes. This is Tom James.

CO-CHAIR GEORGE: Go ahead, Tom.

MEMBER JAMES: I was just -- just to pick up on Tom's point. I think it's important because of the learning curve with all the various electronic medical record systems and the ability to capture data that we start off with eMeasures which are less complex and get that learning curve done as we move then into ones which are more complex, like the ones prior.

CO-CHAIR GEORGE: Tom?

CO-CHAIR KOTTKE: The abstraction of the medical record is not nearly as simple as we think. Because take an academic institution like Duke or the University of Minnesota, we abstracted half the myocardial infarctions, stroke records from 1970 to 1980 when I was a Fellow. And you go in there

and you have an attending physician, you have a third-year resident, you have a first-year resident, and a first-year medical student all reporting something different. What do you take? And so the registers sort of obfuscate these decisions that have been made by a recorder who has made perhaps some arbitrary decisions, so that we're sort of not comparing the same level of complexity.

CO-CHAIR GEORGE: Any other final thoughts before we vote? All right. We'll vote on the measure.

MS. IBRAGIMOVA: Overall Suitability for Endorsement, does the measure meet NQF criteria for endorsement, 1, Yes; 2, No. The results are 94 percent, Yes; 6 percent, No.

CO-CHAIR GEORGE: Would the Committee like to take a break now or go through the next measure before break? If you would like to take a break now, raise your hand. Two? We'll take just a very short maybe five, six minute break.

(Whereupon, the above-entitled matter

went off the record at 10:43 a.m. and resumed at 10:52 a.m.)

CO-CHAIR KOTTKE: Okay. We'd like to get started. Are we ready to get started? Are we -- I see we have our Developers here. All right. Measure 0083, Heart Failure: Beta-Blocker Therapy for Left Ventricular Systolic Dysfunction (LVSD). The developers are AMA-PCPI. The discussants are Mary George and Kristi Mitchell. Developers, do you want to say anything now or just wait?

MS. TIERNEY: I don't think so. Thank you; we appreciate the opportunity.

CO-CHAIR KOTTKE: Okay. Mary?

CO-CHAIR GEORGE: So this is really a paired measure with the one that we just finished discussing, beta-blockers. The evidence is similar to the evidence that we had for the previous two measures today. Level A evidence, 17 randomized control trials, three comparative studies, evidence updated through 2013. I don't have any other comments on it.

MEMBER MITCHELL: The only thing --

CO-CHAIR KOTTKE: Kristi?

MEMBER MITCHELL: -- to add is that they did a QQC.

CO-CHAIR KOTTKE: Okay. Any further discussion or can we vote on the evidence? The evidence sounds like it's -- there's Judd.

MEMBER HOLLANDER: Just one simple question. I'm just kind of curious, this is a one-time prescription of a beta-blocker at any point at either a hospital discharge in the year. And I just -- I'm going to vote for it, but I just wonder whether or not there's not a better measure of looking at it over time, because it seems to me there probably isn't evidence that given a beta-blocker once in a year makes a difference. It's really being on beta-blockers. And I wonder if we're shortchanging what we really want to do by giving people credit for sending them home on a beta-blocker and then stopping it.

MS. SMUK: Yes. So this measure is basically looking at, at a given point in time --- whether it be at a physician office visit or at a

discharge -- that the patient at that time was either actively taking it already, so it wouldn't be that you necessarily have to give it to them if they're already ordered it or already actively taking it, that would qualify, so the provider would just say, I verified that the patient is already -- it's on their active medication list or they already have an order for it or you would, at that time, if that information is not available or if the patient's not on it, then you would have to provide the quality action of actually ordering the medication. And so this measure -- because you can't necessarily look at it at every given point in time, you do have to pick one point in time to look for the quality action.

CO-CHAIR KOTTKE: So it's just a point rather than a period? Yes. Okay. Are we ready to vote on evidence? Evidence sounds high.

MS. IBRAGIMOVA: Importance to Measure and Report, 1A, Evidence Structure Process Intermediate Outcome, 1 High only eligible if QQC submitted, 2 Moderate, 3 Low, 4 Insufficient. So

the results are 88 percent High; 13 percent Moderate.

CO-CHAIR KOTTKE: Okay. Thank you. Opportunity for Improvement, Mary?

CO-CHAIR GEORGE: So the performance gap provided was from the 2010-2013 PQRS data, which showed performance around 75 to 85 percent with really no sustained improvement over the four measurement years. They said that was consistent with the improved HF registry. Disparities have not been noted, but are available at this time.

CO-CHAIR KOTTKE: Kristi, any additional?

MEMBER MITCHELL: Just to reiterate that this gap is reflective of the overall picture and not necessarily around the eMeasure itself.

CO-CHAIR KOTTKE: Great. Any further discussion about Opportunity for Improvement? Hearing none, let's vote.

MS. IBRAGIMOVA: So with the Evidence voting, for some reason, it came out to 101 percent. Do we mind revoting?

CO-CHAIR KOTTKE: Is it round off error?

MS. IBRAGIMOVA: No. It's fine.

CO-CHAIR KOTTKE: Liz says it's a round off error. Yes. Do we have to revote?

MS. IBRAGIMOVA: Just one moment. So Importance to Measure and Report, 1B, Performance Gap, 1 High, 2 Moderate, 3 Low, 4 Insufficient. Just need one more vote. So the results are 24 percent High; 71 percent Moderate; 6 percent Insufficient; 0 percent Low.

CO-CHAIR KOTTKE: Thank you. Specifications and Reliability Testing?

CO-CHAIR GEORGE: So the denominator is patients 18 and older with a diagnosis of heart failure and ejection fraction less than 40 percent, as we heard before, prescribed beta-blocker therapy within the past 12 months in either the in-patient or out-patient setting, limited to the three beta-blocker agents, requiring that the patient has had at least two encounters with the provider in the measurement period.

Exclusions, as we mentioned, are the

standard AMA-PCPI exclusions. And in terms of electronic specifications, all the elements are specified with the VSAC specifications and are specified in the standard HQMF format using the quality data model as required. Heart block was included in the eMeasure specification, but was excluded from the registry specification.

CO-CHAIR KOTTKE: Kristi, any -- Kristi has no further comment. Are we ready to vote on reliability? It sounds like reliability is good.

CO-CHAIR GEORGE: They did do the BONNIE testing for reliability. Now, I guess we cover that under validity testing as well.

MEMBER MITCHELL: I do have a question though. I think in the write-up it said that you all used five different EMR systems. Is that accurate? And if so, can you let us know which five?

MS. TIERNEY: So, I think for the exception analysis, that testing project did include five different EMR systems. I'd have to -- I don't know if we -- I don't think we have it

readily available. We could certainly provide that information. Sorry that it wasn't included.

CO-CHAIR KOTTKE: Okay. We're getting very interesting displays on the screen. So we're ready to -- the committee is ready to vote on reliability. Okay. It's manual labor here. Who's counting?

MS. IBRAGIMOVA: I can count.

CO-CHAIR KOTTKE: Okay.

MS. IBRAGIMOVA: I'll count. Could you just put the slide up for reliability?

CO-CHAIR KOTTKE: High -- yes.

MS. IBRAGIMOVA: Okay. So we're voting on reliability. All of those who are voting High? Zero. Moderate? One, two, three, four, five, six, seven, eight, nine, ten, 11, 12, 13, 14, 15, 16, 17. Seventeen? 100 percent. Okay.

CO-CHAIR KOTTKE: Thank you. Validity?

CO-CHAIR GEORGE: So they did both face validity with an expert panel, as well as the BONNIE testing output. On their face validity testing with 12 responses, eight were agree and four were

strongly agree that the measure could distinguish between good and poor quality. The eMeasure testing found that measure exceptions were validated 95 percent of the time. They looked at 118 exceptions; 98 were for medical reasons for not prescribing, and the overall exception rate was 5 percent.

The data element testing with the BONNIE output used 56 test patients, demonstrated 100 percent performance with 95 percent coverage of the data elements. And the mean performance of the EHR data was 0.9, with a standard deviation of 0.09, which was actually better than their registry testing. This measure is not risk-adjusted. I don't know, Kristi?

CO-CHAIR KOTTKE: Kristi, any additional? Kristi has no additional. Anybody else care to comment on validity? It sounds like validity is good. We're ready to vote on validity.

MS. IBRAGIMOVA: Scientific Acceptability of Measure Properties, 2B, Validity, 1 High, 2 Moderate, 3 Low, 4 Insufficient. So we

have one vote for High, 16 votes for Moderate, zero votes for Low, and zero votes for Insufficient.

CO-CHAIR KOTTKE: Thank you. Feasibility, Mary?

CO-CHAIR GEORGE: So all the data elements were specified, and the BONNIE feasibility scorecard showed that this was feasible for all data elements.

CO-CHAIR KOTTKE: Kristi, any additional? No? Okay. Any other comments on feasibility? Feasibility seems high. Let's vote on feasibility, please.

MS. IBRAGIMOVA: Feasibility -- 1 High, 2 Moderate, 3 Low, 4 Insufficient. Just missing one vote. Just one more time just to capture that last vote. So we have eight votes for High, nine votes for Moderate, zero votes for Low, and zero votes for Insufficient.

CO-CHAIR KOTTKE: Usability and Use, Mary?

CO-CHAIR GEORGE: It's currently used in PQRS, Meaningful Use, Stage Two, and the PINNACLE

Registry.

CO-CHAIR KOTTKE: Kristi, nothing?

MEMBER MITCHELL: The only thing else to add is that MAP in 2014 reviewed the measure and made recommendations for inclusion in the VBPM as well as Physician Compare programs.

CO-CHAIR KOTTKE: Thank you. Any other comments on usability and use? It seems to be used and usable. Let's vote please. I think we have to wait a moment. Okay. Now we can vote.

MS. IBRAGIMOVA: Okay. So Usability and Use -- 1 High, 2 Moderate, 3 Low, 4 Insufficient Information. Just one more vote.

MS. VICALE: Seventeen is the number that we're looking for.

MS. IBRAGIMOVA: So when you're voting, if you can look at your clicker to make sure that the number pops up. If not, your clicker might be dying, so we can switch that one out. So we have nine votes for High, eight votes for Moderate, zero votes for Low, and zero votes for Insufficient Information.

CO-CHAIR KOTTKE: So we're voting on the overall measure -- Heart Failure: Beta-Blocker Therapy for Left Ventricular Systolic Dysfunction -- 0083. It sounds like it is pretty high concordance in all components of the measure. Let's vote on endorsement or not.

MS. IBRAGIMOVA: Overall Suitability for Endorsement, does the measure meet NQF criteria for endorsement, 1 Yes, 2 No. So the results are 17 votes for Yes; zero votes for No.

CO-CHAIR GEORGE: So we'll be moving on to Measure 2740, Patients with Coronary Artery Disease that have a Potentially Avoidable Complication (during the episode time window). Measure Developers? We should note that the next several measures that we're reviewing are composite measures.

MR. DE BRANTES: Well, good morning. And thank you for having us here. My name is Francois de Brantes, and I'm the Executive Director of the Health Care Incentives Improvement Institute. And this is my colleague, Dr. Rastogi.

And we are here to present for your consideration six measures. We originally had seven, but we removed one from consideration, and we'll get into the specifics of why that is.

The measures are very, very similar to one another, so we are going to take a few moments just to give you a very quick overview of what these are and how they're constructed and what our findings have been in getting to this point. Some of these measures, by the way, have previously been endorsed and so we're here back for the endorsement to an extent and then endorsement of new, similar measures.

So these measures are what we refer to as potentially avoidable complications. For the past ten years of work on the development of these measures, we've been very careful to continue to refer to them as potentially avoidable complications because we're not either suggesting or advancing that they're always avoidable, but simply potentially avoidable by a combination of activities from both physicians and the delivery

system generally. If we can advance the slide? All right.

So the measures that we're submitting are the proportion of patients with a particular condition or undergoing a particular procedure that have a potentially avoidable complication during the episode time window. And what we mean by that is we look at these conditions over a period of time. That period of time is what defines the episode. And so the occurrence of an event during that period of time happens to trigger the measure. And so the conditions are coronary artery disease, heart failure, hypertension, arrhythmias, and then patients undergoing an angioplasty or undergoing a pacemaker or defibrillator implantation. Next slide.

So how we calculate these rates of avoidable complications are by looking at, as a denominator, patients who have that particular condition and those conditions, again, are defined as episodes. All of our definitions for these episodes are posted as open source definitions on

our website, so anyone can use them, anyone can look at them, and anyone can also give us feedback on their definitions. The numerator is defined by the number of these patients with a potentially avoidable complication during the episode time window. Next slide.

There are two types of potentially avoidable complications and how we define these, Type Ones and Type Twos. And all that is in the materials that we've submitted. The Type One complications are complications that are very directly related to the condition or the procedure. Such, for example, as an emergency department visit for a patient, and in this case we're using hyper or hypoglycemia in diabetic patients. We're not submitted diabetes, but obviously the same kind of rationale applies for patients with CAD or heart failure. And these are complications, as I said, that typically are best controlled by the managing provider.

And then the Type Two avoidable complications are those that are related generally

to patient safety failures and include, for example, the CMS-defined hospital acquired conditions. And these are avoidable complications that generally are best controlled by process improvement beyond just the managing physician. Next slide.

So these are the results of our testing on these measures. And what we're showing here is a risk-standardized PAC rate, or rate of potentially avoidable complications, for each one of the submitted measures, plus one AMI that we're actually not submitting. And, again, I'll get into that in a minute. So this shows you the distribution of the rates of avoidable complications and their spread on the particular data set on which we studied these complications. Next slide.

So we spent a tremendous amount of time looking at and evaluating the reliability of each one of these measures. And our conclusions on what constitutes a reliable measure is first of all whether or not it can distinguish provider

performance. And it can distinguish provider performance -- a measure is reliable if it can show that there are some high provider-to-provider differences and then within provider variability it ends up being relatively low.

What we found is that reliability, of course, is a function of whether or not there is a cross-provider variability of the Measure. The less variability, the less reliable the measure is because it just simply doesn't differentiate performance from one to another. The second element is the number of comparative providers. If you have very few providers that you're comparing, then of course it's relatively difficult to have a reliable test result. And the third one is the sample size. As we all know, typically the lower the sample size, the less reliability the result ends up by being.

The criteria for achieving reliability and our conclusion on the criteria for achieving reliability varies not simply from episode to episode, but also from data set to data set. And

that's an important point that we wanted to make because the data sets that we worked on to assess the reliability of these measures are commercial data sets. So we are not inferring in any way that the results of the reliability testing would be the same if we tested these measures on say a Medicaid data set or a Medicare data set. All right. So we're very specific that what we tested these measures on are commercial data sets, and we know that there are differences in the reliability scores when we move from one data set to another data set.

MEMBER DELONG: Could I ask a question? When you say episode, do you mean one of these areas of preventable complications? I mean, you're not talking about episodes of care; you're talking about these categories.

MR. DE BRANTES: Right.

MEMBER DELONG: Is that correct?

MR. DE BRANTES: So just to be clear, if you look at this slide, so the episodes are, or what we define as an episode of CAD, is actually 12

months of -- looking at the 12 months' worth of claims data for the management of a patient with CAD, 12 months' worth of looking at the management of patient with heart failure. Same thing with hypertension, arrhythmia, heart block. PCI is a shorter term episode; we look at 90 days post-procedure as the time window for assessing whether or not there was an avoidable complication. And same thing for pacemakers and defibrillators. So we're looking at a period of time, a time window, and whether or not there were avoidable complications during that time window, and only during that time window.

MEMBER DELONG: So it's more like an episode of care rather than the categories you have? I just didn't understand.

MR. DE BRANTES: Yes, that's correct. So what our results, again, in the data sets that we used -- which are commercial insurer data sets of millions and millions of plan members, so it's not a small data set by any stretch of the imagination -- is that we achieved relatively high

reliability with relatively low sample size for most of the condition episodes. This is directly a function of the very high degree of variability that you can observe in rates of avoidable complication from provider to provider who manage patients with those types of conditions.

For procedures like PCIs, it's a higher threshold, so you need a higher sample size. And for pacemaker, defibrillator, we used two different data sets. And you can see the difference that the data set makes, because in the first data set, we were able to achieve an absolute reliability with a sample size of 128, and then for the next data set --- the other data set that we used -- the sample size went down to 22. So, again, it's very important to understand that these reliability rates change data set by data set.

And one of the continuing points that we made to our colleagues at NQF is that we want to be very specific about that, because we don't want these measures to be used by anyone inferring that because you have reliability in one data set that,

that reliability automatically confers to any other data set, right? So, you have to be rigorous about how you apply measures and not make inferences of reliability when they're not appropriate.

MEMBER DELONG: Could I ask another question?

MR. DE BRANTES: Sure.

MEMBER DELONG: When you're talking about reliability --

MR. DE BRANTES: Yes.

MEMBER DELONG: -- and sample size, it seems that you're really talking about whether you have a significant p-value and not necessarily the range you're seeing in the variability. I mean, it's possible that with five cites, you're seeing the same range, but you don't have enough power to claim a significant result. That with five cites, you've got the same range as with 128 cites in terms of the hospital-specific --

MR. DE BRANTES: Yes. It's 128 patients, by the way.

MEMBER DELONG: Well, yes. Either way.

DR. RASTOGI: So, you're right, in a way that variability in that data set is the one that drives the reliability score. So if there's no variability across providers, then the beta binomial model shows the alpha and beta values to be very different. And if there's high variability in that data set, then we found very different numbers there. So you're right, I don't know if it's a p-value, but it's the across-provider variability calculation. If it's 0.0005, then there's no point calculating a reliability score. So that 128, the star is there that even though we derived a number, but that across-provider variability was not significant.

MR. DE BRANTES: And that's one of the reasons why --

DR. RASTOGI: So, in the second --

MR. DE BRANTES: Go ahead. Yes.

DR. RASTOGI: Sorry. In the second data set, it was very significant, and so we could come up with the number 22 without asterisks around it. And they were two different commercial data sets.

So it was very interesting for us, and that's where we are saying that when this measure is applied, then first they have to check the reliability scores. And only give a performance number for a physician if they meet that minimum sample size for which --

MR. DE BRANTES: Right.

DR. RASTOGI: -- say the reliability is over 0.7.

MR. DE BRANTES: Exactly.

MEMBER DELONG: So I mentioned this yesterday: sometimes when you produce parameter estimates, they're not as meaningful as seeing the actual numbers. For example, what does that translate to in a risk-adjusted rate that we can understand and the variability in those rates? It would be helpful to see that range.

DR. RASTOGI: That's right. And our workbooks provide that level of detail.

MR. DE BRANTES: And we had the summary in the prior slide where we showed the risk-adjusted, risk-standardized rate of

potentially avoidable complications. And this is one of the reasons why we also removed from consideration and, in fact, let lapse a prior endorsed Measure around rates of potentially avoidable complications for patients with an MI, because we simply were not able to get enough of a reliable score -- at least to our standards -- on that particular measure with any commercial data set that we were analyzing.

So it's not to say that we wouldn't be able to get a reliable score with a Medicare data set, and we actually have some -- we've run some internal tests that would suggest that we would in fact get a reliable score on a Medicare data set. But we're not here to get this measure endorsed for Medicare; we're here to look for endorsement for commercial populations, not Medicare populations.

DR. RASTOGI: Yes. And just to add to Francois's point on AMI, it was very interesting because in the commercial population the AMI rates have really dropped, even in the last few years we are watching. And because the rates are so small,

the highest sample size was only 45 patients for a facility where the AMI got attributed. And so, it was -- the beta binomial just didn't give us good numbers.

MR. DE BRANTES: And I think it shows that the good work that other measures are doing in improving care of patients who have heart conditions and as a result of which, there are fewer MIs for commercial populations, and that's good news for all of us. So why have a measure that doesn't create differentiation of providers for something that's not necessary? And that's why we decided to take it off consideration. Next slide.

So here's an example of, and I think it goes to the question you just asked, here we're showing the results for five different providers of their risk-standardized rates of avoidable complications. And what was interesting here for us is that it again points to the fact that making broad inferences about provider performance based on one single measure is probably not a good idea. And while we all know that, I think this slide makes

the point.

Because the reality is that when you look at the distribution here, apart from Provider E, who seems to be certainly worse than the average on pretty much all of these measures, for the other ones, it's a mixed distribution. And certainly we could probably say Provider C is average to better than average on almost all of these specific episodes, but for the others, it's a mixed picture.

So I think, again, the point is while the measures are similar, they show different results. And it's important for measure users to not make inferences about the performance of a physician or a hospital based on the results of one measure, say the Management of Patients with CAD or the Management of Patients with Hypertension. But instead, be very clear about the fact that there is variability. And that the results of that variability can show good performance on one measure and not show good performance on another measure.

So in summary, I think what we're

submitting and what we've shown in the documentation submitted is that rates of potentially avoidable complications, adjusted for patient severity, can reliably differentiate provider performance, that minimal sample sizes to achieve scoring reliability vary episode by episode and data set by data set, but that the performance on rates of avoidable complications also varies by episode and by provider, even when you're looking at the management of chronic conditions in the same clinical domain.

MEMBER AL-KHATIB: So thank you. I'll be actually one of the reviewers of the pacemaker, defibrillator measure. And I know I can delve into the details then, but I would like to bring up a couple of concerns because I think those apply to all the measures that we're discussing today. The definition of the PAC for me is difficult to wrap my mind around, if you will. And the main reason being that I think it's vague, I think it's very broad.

Especially when you look at the second

type, like the first type that you showed, yes, that's directly related to the device implantation. I can buy that. The second one is way too vague, way too broad. And I think that applies to all the measures. I don't know how the other reviewers of the other measures feel, but I would like us to be consistent as we review these specifications.

MR. DE BRANTES: All right. So if I can respond? You're right, they are vague. Well, I don't think they're vague, they're actually quite specific, but they're broad. And there's a reason for that. It's because ultimately, it's about the patient. And these events occur, and we're measuring the frequency with which they occur. They occur to patients and in a world that I think we all agree is moving away from individual accountability to team accountability, even if that's a difficult concept still for many physicians to accept, that's really what this is about, right?

The PAC Two Types are about joint accountability of facilities and managing

physicians and others involved in the care of the patient and how they're collectively working to try to minimize these types of patient safety failures. And it's -- if you think about the value-based payment world, for example, CMS has just announced their mandated comprehensive care joint replacement episode. It's going to include all costs post-discharge for 90 days. I can tell you that every single one of our potentially avoidable complications is something that the hospitals will be fully accountable for.

And so that's where the world is moving towards. And think about ACOs, right? I mean, if you're in an ACO, every single one of these avoidable complication applies. And you might say, well, it applies to the ACO, that's fine. But, isn't the ACO the collective of all of the providers that make it up? So, ultimately, what we're trying to do here is to say, we're moving to a team sport and these rates of avoidable complications vary tremendously provider by provider, and can, therefore, help them really understand what's going

on with these patients both when it's within the direct control of, say, the surgeon and also when it's not in the direct control of the surgeon, but in the control of other team members that are co-participating in managing that patient's episode.

MEMBER AL-KHATIB: Well, I actually have major concerns about calling these safety failures. I mean, while some of them may be related to something that the physician or somebody on their team didn't do, there are a lot of factors here at play. You have --

MR. DE BRANTES: Sure.

MEMBER AL-KHATIB: -- the patient perspective too. You send them out on medications, they don't take the medications, they come back. But are you saying that if they actually have to come back for any reason during that time, that the physician who did the procedure will be dinged for it? I mean, I have major concerns --

MR. DE BRANTES: Well, isn't --

MEMBER AL-KHATIB: -- about that.

MR. DE BRANTES: Isn't that the way the all cause readmissions work as well?

MEMBER AL-KHATIB: I'm just --

MR. DE BRANTES: No, no, no. But --

MEMBER AL-KHATIB: -- talking about these particular measures.

MR. DE BRANTES: -- I'm saying, isn't that the way the all cause readmissions work --

MEMBER AL-KHATIB: I do not --

MR. DE BRANTES: -- as well?

MEMBER AL-KHATIB: -- know. I don't know if the person comes back with a stubbed toe, is that considered a failure?

MR. DE BRANTES: All cause readmissions is all cause readmissions.

DR. RASTOGI: So let me add to that. Being a physician and a cardiothoracic surgeon, and I appreciate your sentiment, but as a cardiothoracic surgeon, I want to emphasize that I want to be accountable for my patient no matter what happens to them. The stubbed toe is outside the episode.

MR. DE BRANTES: Right.

DR. RASTOGI: So when we create the episode, all kinds of accidents and things that are not related to the episode are not included. So it's not all cause readmission, which is even broader than that.

MR. DE BRANTES: Right.

DR. RASTOGI: It is specific to the episode. But then the patient safety failures could be process failures, say central line infections, et cetera, that happen in the hospital. And if I'm operating in two different hospitals and if in one hospital my infection rate is higher and in the other one it's lower, I will move my practice to the lower infection rate. Because I don't want my pacemakers to get infected, whether it is because of me or it's because of a safety failure because of poor practices at the hospital.

MEMBER AL-KHATIB: But the example that you're using, there's a direct link. I mean, absolutely, if they come back for an infection for any reason, I would have to say, did I contribute

to that? But if it's completely separate from what I did with the procedure -- let me give you an example. What if the patient receives and ICD and they come back with an ICD shock for a VT that they had? Is that -- I helped them, I saved their life. Are you telling me that, no I'll be dinged because they came back for an appropriate ICD shock?

DR. RASTOGI: So we have defined these complications specifically. The ICD shock is not there, okay? So if you look at the list of PACs, they are all given. This is not as part of the list.

MR. DE BRANTES: Right.

DR. RASTOGI: So ICD shock is not there.

CO-CHAIR GEORGE: I'm going to cut this off. I think these will fall into the measures specifically, and it's a great discussion. I can see there's a lot of concern, but I think we'll move on to the first measure. Leslie?

MEMBER CHO: Great; thank you. Okay. So, we are going to talk about 2740, which is Proportion of Patients with Coronary Artery Disease that have a Potentially Avoidable Complication

(during the episode time). And the complications are for 12 months; is the correct? So I actually looked through your potential avoidable complication list, which has 789 potential avoidable complications. And in some of those things, I totally understand the reasoning. In others, it boggles the mind why they are included. And I will give you some of my favorite examples from the 789. Here we go.

My favorite examples are fall from a wheelchair or fall from a bed, fever which is unrelated to any kind of catheter insertion or anything like that. I think my number one problem is: I totally fully support the idea in which this measure was developed, which was that you want to hold the physicians accountable for what has happened in the hospital.

What I don't think we should, and there is no evidence for, is the one year time window. And the 789 potentially avoidable complication is -- I find it unfathomable that anybody would hold a physician, even if we put that potential word in

there, responsible for these. I just -- it boggles the mind.

DR. RASTOGI: Okay. So let me respond to the 749 first or 789 or as many number. Our measures are based on claims data, okay? This is administrative claims data. So I understand quite a bit of the numbers or the rules are coming because of individual codes, all of them may map to the same thing, say line sepsis, it's not one code, but there are several codes. So when you really count the types of complications that the numbers, you can group them all and say if any of these codes come in, it suggests it's line sepsis. And then that complication is really line sepsis.

MEMBER CHO: But how does that make any sense? So I would understand -- that, to me, makes no sense to me. Because what it tells me is, if I have a patient that got admitted for unstable angina, which is CAD, which is one of the things that will get a patient in here, and then they go home, I treat them well, they go home and in a couple of weeks later or months later, they fall out of bed

because they tripped on a rug, then it is counted as potentially avoidable complication.

MR. DE BRANTES: Right.

DR. RASTOGI: No, only if that claim carries a CAD diagnosis.

MEMBER CHO: But it will -- so we are both, you are a cardiothoracic surgeon, I'm an interventional cardiologist, we do DRG coding, we do ICD-9, now we're going to do ICD-10, God bless, and the --

(Laughter.)

MR. DE BRANTES: Or not.

MEMBER CHO: -- but the thing is, when a patient comes in and we list the discharge diagnosis that somebody will put in fall out of bed, fracture a femur, and then put CAD as a secondary diagnosis.

DR. RASTOGI: So the principle --

MR. DE BRANTES: It wouldn't count.

DR. RASTOGI: -- diagnosis counts, the secondary doesn't count, for the in-patient stay. So each claim goes code-by-code and there are rules and definitions, and they're not going to be

unreasonable to physicians, okay? The idea of creating these measures is not to hold physicians -- I'm a physician, I want to protect them honestly. So I was very, very careful when we entered these definitions and the rules.

MEMBER CHO: But help me understand then.

DR. RASTOGI: Yes.

MEMBER CHO: Because I just don't understand it then. So help me to -- explain it to me.

MR. DE BRANTES: All right. So let me address one issue, which is when you look at the data, what you see is that these what you might consider random events end up by being distributed fairly uniformly across all providers. So it's not -- that's not what creates the signal in the overall rate of avoidable complications because it's noise in the data. But sometimes the noise can get to a point where it actually creates a strong signal. And that's the point which is -- you want to know what's going on systemically with the patients.

And the noise everyone ignores, but when

there is a signal strength, then you can pay attention to where that signal strength comes from. And we can't predict where that signal strength comes from. But we can observe where it comes from when we run the analyses in data sets. And that's what's important, is the feedback about the signal strength of the measure and what it tells you as a practicing physician about what's going on with your patient, both in an in-patient setting as well as an out-patient setting. The randomness of the patient who might have an event here and there is just not going to create signal strength in the kinds of reliability tests that were shown.

MEMBER DELONG: If you have enough noise in your data, it will camouflage the signal.

MR. DE BRANTES: But it doesn't appear to, right? So I mean, our testing is pretty clear on that.

MEMBER VIDOVICH: So, I'm on the review for the hypertension measure, and I had very similar concerns that Leslie had. And I feel this is some sort of a big-data measure that I see like, and I'll

give you an example that's been bothering me. Several years ago, there's an article in New York Times how insurance companies look at the purchase of generic car change oil in a huge department store and then predict default on credit card debt later on, right? Assuming you don't have enough money to go to, whatever, Midas, right, and you buy your oil and change your own oil. So they included a huge number that a big data credit companies can include.

And, again, looking at some of the level one potential avoidable complications, let's say for hypertension, VF, acute systolic heart failure, I would buy that, right? But that's something I could look in a year later. But then, there's another one, level one, how are they causing adverse effect with therapeutic use? I mean, that's really tough to put together a general anesthetic not being used. And then if you go further down is antipyretics causing adverse effect with therapeutic use as a Type One potentially avoidable complication. If you go further down in this list, then it's impaction of intestine unspecified,

that's a level two. And I can go on and on, there's a bunch of them. Poisoning by vitamins not elsewhere classified.

So certainly, when you throw in a lot of data and you throw in -- there's a statistical model that they're using, this is a, I think, regression model, you will end up with some sort of statistical, but this is association. This is by no means causation. And the more numbers you put in there, you will get a statistical, and if it gets up to 0.9 the more episodes and the more provider you include. But I seriously doubt that we can do big data association to ding a physician for some potential causation. I don't know. It might be very -- there may be huge unintended consequences to this that I see here.

CO-CHAIR GEORGE: Sana, Linda, and Judd?

MEMBER AL-KHATIB: The one comment that I want to add is I appreciate your clarification that what you're looking at is -- you're not looking at people with none of these complications. You're basically looking at outliers, if you will, people

who have worse outcomes than other people. I appreciate that totally.

But that also is a reflection of the patient case mix that you have. And you're going to say, well, we adjusted for that. But at least for the pacemaker and ICD variable, I mean, measure, and we're going to get to that, you use 170 variables in the model to adjust for those. That to me creates a lot of concerns. I mean, is it practicable? Is it even doable to think that we're going to able to adjust for 170 factors? And even after the most rigorous adjustments, can you really adjust for and eliminate all those differences in the patient mix? So that is another unintended consequence that people may start now selecting patients. That I'm just going to operate or implant pacemakers, ICDs, on the healthiest of the healthiest.

CO-CHAIR GEORGE: So I just want to mention that we're still discussing the evidence. Go ahead.

MEMBER BRIGGS: So I would say that all

of these measures kind of have similar problems. And on -- I have the heart failure one. And so one of the things that I was looking at the 800 plus untoward things that can happen, PACs, in those groups of patients. And an unforeseen complication of using this measure, I think you're going to be dinging a lot of people for hypopotassemia is one of the untoward things that can happen.

So if a diligent provider is monitoring their patient and getting blood work to check and make sure that the diuretics are not causing problems there, then you end up with a diagnosis of hypopotassemia or hypokalemia. And if you're going to treat that, you want to be able to bill for it, so you're going to put the code down, which means now you've gotten an untoward complication for something that you're trying to avoid. That you've actually monitored for it to try and prevent it, and now you've found it -- even if it's a K of 3.4 and you decide you want to give them K-dur for it or tell them to do something like drink orange juice or

whatever, change to a potassium-sparing diuretic -- same kind of thing.

You're actually hurting the provider for monitoring to avoid complications in that particular scenario. And you picked that up because it's associated with heart failure. You see that and hyponatremia, all those things, when you do that data analysis that you're talking about.

CO-CHAIR GEORGE: So, Judd, new comments?

MEMBER HOLLANDER: I mean some of these aren't potentially avoidable, I mean, these are just things that are going to happen. And some of them are going to happen because it's a side effect of the treatment that is evidence-based, based on the guidelines and you check it because you expect it to happen. I don't think when we start somebody on a diuretic, we expect their potassium to be normal, particularly at the beginning. We monitor to see where it's going to go. So to me, that's not even a potentially avoidable complication; that's good monitoring and detecting something that we

expect with happen some percentage of the time.

I think my bigger issue with this is: what's the right answer? It shouldn't be zero. What is -- even if you accepted there's 890 things that are all preventable and we could do that, where should we be living on this? That's one. And then the final thing, which I think is more important, hypokalemia and death are not on the same scale, but if they all add up and just count as a one, that's a little bit of a problem to me. Because there's preventable catastrophic things and there's relatively trivial things.

CO-CHAIR GEORGE: New comments that haven't already been discussed?

MEMBER VIDOVICH: Just one quick comment. For the adjustment, a multiple adjustment is done for baseline medical conditions, but there is no adjustment for race, at least in the hypertension, nor sociodemographic features. So you may actually -- so I think this is incompletely adjusted data, the way I see this. At least the hypertension measure.

CO-CHAIR GEORGE: Any new comments on the evidence that we haven't discussed?

MR. DE BRANTES: On the last point, there is no way to adjust for socioeconomic difference using claims data. So while we wish there were, there isn't. And the moment that health plans and other payers systematically capture the ethnicity of the patients, then we can include those and adjust with models. But until then, you can't do it.

MEMBER DELONG: So this is something that really concerns me. Because we are going to assess performance of providers and we keep saying, we don't have what we really need, but this is the best we can do. Well, I would not want to be assessed on the best we can do if it is not adequate.

DR. RASTOGI: Maybe I'll comment on that. There was a very interesting study that New York Montefiore Hospital took on using our measures because they were very interested in our PAC measures. And they took great pains to do a sociodemographic adjustment. So they found the

race, ethnicity; they matched it up; they linked it to claims data. One physician put his whole life on it. How many years? Three, four years ago.

And they took 18 months trying to get that data matched up, analyzed, and in the end after they did the risk-adjustment, they realized that this was all pointless. And the reason is, they were increasing the disparities, and they realized that doing this thing -- this kind of adjustment was sending the wrong signal to the physicians in New York State, and so they decided to back off. The guy quit his job.

MR. DE BRANTES: Right. The point being, if you adjust for -- in their instance, their conclusion was, if you adjusted for the ethnicity of the patients, you would perpetuate a difference in, accepting a difference in the treatment of patients of different races.

MS. MARINELARENA: So again, back to evidence. I just want to remind everybody what the requirement is for an outcome. So we're looking for the rationale supports the relationship of the

health outcome, outcome to process, or structures of care. Okay. I don't think we've had a discussion on the evidence itself.

MEMBER CHO: Well, I think we made our -- at least for me, I think there is no evidence that the rationale supports the relationship of health outcome. Because for the 789 potentially avoidable complication, in my mind, I just don't see the evidence.

MR. DE BRANTES: Right. Let me just make one final point on that issue. As payment moves, whether people like it or not -- we happen to like it -- but as payment moves to value-based payment, all this is in. Whether you like it or not. All this is in, and then some. And then some.

MEMBER CHO: I'm only one person voting my conscience on this thing. And all I have to say is: this is a measure, my name is on one of these committee things, and in my honest opinion, I cannot endorse a measure that holds a physician responsible for one year for potentially avoidable complications that are 789. I cannot. And

there's no evidence in my mind. And I mean, that's it. And I understand the field going this way; I understand the insurance companies; I understand the state of the healthcare in America. But I just have to vote how I feel.

DR. RASTOGI: I want to make a couple of comments, just feedback on the comments that were made. Death is not a potentially avoidable complication in our measure, partly because it was very difficult to measure death. Once they are dead, then their claims didn't come back, unless it was an inpatient mortality, so mortality's not there at all.

These are PACs that perhaps could be avoided by the physician or by the team that's working with them. I appreciate your point that one PAC could be very minor, and another PAC could be very big, but for a provider, if they have one bad outcome amongst all their patients, it's just one bad outcome.

But if every patient has some bad outcome or the other, then there's something going

on. There's some issue with that physician. As Spencer mentioned, we are seeing this variability. So when you talked about fluid and electrolyte disturbances with hypokalemia as an example, because it was in the comments, I reviewed it back.

Only 5.79 percent of the heart failure patients had that PAC listed there, so only 349 out of the 6,000 odd episodes had that particular issue. If heart failure hypokalemia is such a big issue, it didn't show up in our case. Just to make that point that whatever we are seeing in the data, that's what is being captured.

So when you said that some of these PACs don't make sense, yes, some of it is also derived because the PACs were done based on the clinical classification software from EHRQ. So if it's written poisoning, much of that whole group has been taken together. To the extent it doesn't happen in a patient, it never shows up, and the baseline is what we are looking at, where is the baseline, and then is their signal -- is a given provider significantly worse than that baseline, and can we

keep pushing this baseline down? Can this provide transparency? Can this provide a way to help providers improve?

Since our measures have been used in that way in many states, including like Calcas in California use this for process improvement. They were so excited. The physician group took it on, and they wanted to test it, and they wanted to do process improvements.

CO-CHAIR GEORGE: Excuse me. We have a question on the phone from one of our members.

MEMBER JAMES: Yes, hi. This is Tom James. To jump in on this, representing a Medicaid plan recently, and now with a commercial health plan, I think we do have a lot of that access to information. If we go back to -- so we can understand the ethnicities and the race, the whole class set of standards.

If we go back and look at what is our whole point, which is improving the healthcare of this country, and we use this not with the expectation that we will ever hit 100 percent, but

that we can look to constant improvement over time on this kind of a measure, I think this is an important measure, and it is a positive one. Thank you.

CO-CHAIR GEORGE: Thanks, Tom.

DR. BURSTIN: Just one quick process point again. I know this is difficult. Two quick things. This is not at the individual physician level of analysis. I think that's really important. When these measures came through the first round to our prior committees, they felt strongly this would not be appropriate at the individual level.

In fact, this has come back at the group and health plan level, so just one point of clarification because we keep saying the individual doc, the individual doc. It's not. For many of us who practice in academic health centers, the group is enormous. So again, keep that in mind.

Secondly, we want you to -- obviously, we're here because of your expertise, but we also need you to look at what's on the forms. So this

is really about is there evidence of the rationale? So again, there is a rationale provided. You can accept it or not accept it, but it's a little bit less about your conscience and more about looking at our criteria and looking at what's on the form. I just want to be just a little more clear on that.

MEMBER CHO: So in the worksheet, it actually says clinician. Level of analysis is clinician, page 1.

DR. BURSTIN: It's clinician slash -- it says group.

MEMBER CHO: And then clinician colon individual.

DR. RASTOGI: That's right.

DR. BURSTIN: But that's not the case anymore, sorry.

MR. DE BRANTES: But again, it's contingent on the testing or the reliability of the measure, and it's at multiple levels of units of measurement.

If you don't get reliability at the individual provider level, then you look at the

practice, and you look at the group, and you look at the health system. The goal is to get to reliable measures on individual physicians. There's no question about that. That's what we all want. That's what consumers want. Yes, absolutely.

DR. RASTOGI: But quite often, providers work in groups, so if there's coverage over the weekend and they would rather submit it together, as a team, then that option exists.

CO-CHAIR GEORGE: George?

(No audible response.)

CO-CHAIR GEORGE: Yes, I would just say that I'm trying to understand how these things are related to something that the physician/physician group practice could make actionable. I think for me, there's just a lot of disconnect. That's just my opinion. Any other --

MEMBER PHILIPPIDES: -- before, but that's sort of what I was struggling with. Before we vote, is the level of analysis here at the physician level? Because it's unclear from the way

that it was written. You're saying, no, Helen, it's not. It's at the facility/hospital level?

MR. DE BRANTES: It's at the level --

MEMBER PHILIPPIDES: For me, that's a big difference before I vote.

MR. DE BRANTES: It's at the level at which you achieve reliability in the score of the measure. If you can get it at the physician level, then it's at the physician level.

DR. BURSTIN: The form says (Simultaneous speaking).

MEMBER BECKER: I want to make a point. We're discussing evidence. The evidence that's been supplied is that there's an association between these things and a diagnosis.

An association is not causality. We have no systematic reviews. We have no randomized controlled trials. We have nothing that actually says that those 792 or the 800 have any causality in any of this, and we don't have the kind of scientific data that we've been talking about for evidence review.

MS. WILBON: Because this is an outcome measure, that requirement is different, obviously, because process --

MEMBER BECKER: Still, association is not causality.

(Simultaneous speaking.)

MEMBER VIDOVICH: It's really -- I have to say, really bothering me. Since there is no causation, what is the physician to do? Because we know that ACE and ARB improve every function in heart failure in people to reduce mortality, but what is a physician to do to reduce one of these 750? I don't see the causality. What will either a physician or healthcare system or the intake nurse do to improve this number, which is not zero or 100?

MR. DE BRANTES: If I could answer that because the measures are actually in use in multiple parts of the country. What the physicians have done, first of all, is generally thank us for the reports that are generated on these avoidable complications and immediately keyed in on potential issues that get to causality.

Claims data don't get you to causality. We're not inferring causality. What we're observing is that these events are occurring, and they're occurring at a high frequency and very variably.

When the physicians get the reports back, they've systematically gone back to medical records, looked at, and tried to understand the potential causality, and then acted on that and reduced their rates of avoidable complications. So these measures actually happen to be very actionable, and those who have gotten the reports on them have found them very useful.

MEMBER VIDOVICH: Do you have data to demonstrate that there is actually -- if you act upon these measures, you get results?

MR. DE BRANTES: Yes, we've seen rates of avoidable complications go down in all of our implementations.

DR. RASTOGI: Yes, and interestingly, like the heart failure rate for the inpatient, it's only 5 percent now. Six years ago, seven years ago,

the whole CHF PAC rate was more than 70 percent, 75 percent.

CO-CHAIR GEORGE: Okay, let's really stick to the evidence question. Any more comments on the evidence? All right, we'll vote on the evidence.

MS. IBRAGIMOVA: Importance to measure and report 1A evidence, health outcome or PRO, 1 yes, 2 no.

(Voting.)

MS. IBRAGIMOVA: It's capturing. It just doesn't say it on the screen. The results are 3 votes for yes, 14 votes for no, so it does not pass.

CO-CHAIR KOTTKE: Okay, 2747, proportion of patients with heart failure that have potentially avoidable complication during the episode time window, Michael Crouch and Linda Briggs. Linda's going to do the primary.

MEMBER BRIGGS: I've kind of already said my piece about the evidence related to the heart failure. It kind of goes the same way that just because you run the dataset, look at the

frequency of things that happen with people that have heart failure, that association does not mean that there is a link there. Yes, there could be, but we're asked, again, to look at the quality, quantity and consistency of the evidence, and it's not there.

CO-CHAIR KOTTKE: No, you don't need to look at QQC for outcome measures.

PARTICIPANT: Right.

MEMBER BRIGGS: Okay.

CO-CHAIR KOTTKE: It's just is there a potential association -- or an association, sorry.

MEMBER BRIGGS: Still, the evidence that was presented was tangential. It was not directly related to the measure. Michael, do you have anything else that you wanted to say about that?

MEMBER CROUCH: If the list of heart failure related potentially preventable things were longer and more clearly potentially related to heart failure, I'd be all for it. However, including things like psychotic break, one of my

favorites, how can you possibly relate that to heart failure, I cannot conceive.

There's a long -- gastritis. How that's related to heart failure, beyond me. There's way too many things in here that even if they're specifically associated with patients, with certain providers having more of those things, what that has to do with heart failure is totally obscure to me.

CO-CHAIR KOTTKE: I'm sensing a direction. Judd, do we --

MEMBER HOLLANDER: A different comment on this. One of the things that perplexes me and makes this hard for me to get over is I believe this is the first set of measures I've seen in my three times around this table where I don't really know who I'm measuring.

I mean I understand your answer. It's where you achieve reliability. I actually would be totally fine with this at a health system level or an ACO level because they're supposed to manage care of the totality of the patient. I'm not okay with

this at the individual clinician level for the reasons that people have voiced.

We have seen measures come in, I don't remember the numbers, but Measure 1, which is blah, blah, blah at the clinician level, and Measure 2 is blah, blah, blah at the health system level. I think this might actually be easier to accept, at least by me, if it was at a bigger level, rather than potentially at the individual clinician level. I'm trying to be at least a little bit encouraging because there are things that make sense, but it might not to the group -- the individual clinician level.

CO-CHAIR KOTTKE: Ellen.

MEMBER HILLEGAS: I think this would be good discussion for a minute or two of what would be acceptable? Because I think, also, cutting down the number from 769 with some of these ridiculous -- 846 this time, I'm sorry. If we cut it down to a reasonable number of appropriate complications, I think that might also help drive all of these measures forward.

I think these measures are very, very valuable, but it's very similar to what we see in the hospital, where they've decided there are too many infections with indwelling catheters, so everybody's catheter gets pulled out, even if they're 24/7 bed rest on propofol.

So some of these things are a little bit ridiculous. Maybe if we cut down the number and made it appropriate for the diagnosis, as well as made it at a healthcare provider or facility level, and not a provider, I think you could have a very, very strong measure. But I think at this point, from what we're all talking about, the number of the problems and the level that you're looking at, I don't think any of these measures are going to go forward.

But I do think they're valuable. I really do think the information is valuable. I think you have something here. I agree with you that we're going to value-based, but I think at this point, right now, what you've presented us probably is not going to pass because of the number and the

level that it's at.

PARTICIPANT: Yes, no I mean we -- I think that message is pretty clear.

CO-CHAIR KOTTKE: Let me ask a rhetorical question, here. Does anybody feel that they're going to vote markedly different than they did on the last measure?

Then I think we ought to just vote and move on. Does anybody object to that strategy? We're voting on the evidence.

MS. WILBON: Oh, okay. Yes, sorry, I thought you were --

MS. IBRAGIMOVA: Importance to measure and report 1A evidence, health outcome or PRO, 1 yes, 2 no.

(Voting.)

MS. IBRAGIMOVA: Tom, can you please send your vote via text?

MEMBER JAMES: Oh, I did vote text and chat this time.

MS. IBRAGIMOVA: Okay, thank you. We just received it.

MEMBER JAMES: I just wanted to get double credit.

MS. IBRAGIMOVA: The results are 2 votes for yes, 15 votes for no. This measure does not pass.

CO-CHAIR KOTTKE: Okay, I've been instructed by Mary to do the same thing for 2748, proportion of patients with hypertension that have a potentially avoidable complication during the episode time window. Mladen? Henry's somewhere outside the country.

MEMBER VIDOVICH: I think this is a similar measure to the ones we previously discussed, and I think much has been said. I would say I'd like to take the opportunity to say I think this is the future. I think big data is the future, and I think increasing computing power will allow us to do this, but I think these are not ready for prime time.

I think they need additional adjustment, additional refinement, and I think with some, maybe, prospective data to show us that really

acting upon this number does change outcomes and benefits the overall population. But at this time, as presented, I don't feel I'm comfortable voting for this.

CO-CHAIR KOTTKE: Mary wants to say something.

CO-CHAIR GEORGE: I just wanted to ask if the evidence was the same for this measure as it was for the first measure?

MEMBER VIDOVICH: Yes, there is also similar absence of evidence, yes.

CO-CHAIR GEORGE: Okay.

CO-CHAIR KOTTKE: I'm sort of reminded of a paper I wrote and looking at when Oliver Wendell Holmes suggested that obstetricians spread puerperal sepsis, and all the obstetricians said we don't understand how the hell this can happen. How can you blame obstetricians for these things that obviously have nothing to do with obstetrician behavior? Now I'll catch the slings and arrows of outrageous fortune from you guys, but I think there are associations here. Are we ready to vote?

Anybody have --

MS. IBRAGIMOVA: Importance to measure and report 1A evidence, health outcome, or PRO, 1 yes, 2 no.

(Voting.)

CO-CHAIR KOTTKE: While we're waiting for the vote, I'd also like to raise sort of a rhetorical question. If they came back at a facility level, sort of show of hands, how many people would say that they'd consider that this would be something that could --

MR. DE BRANTES: We're not doing that.

CO-CHAIR KOTTKE: Okay.

PARTICIPANT: Updating facility for --

(Simultaneous speaking.)

MR. DE BRANTES: No, I know, but we're not going to change all of our measures that are important to consumers to please a committee.

CO-CHAIR KOTTKE: Okay, that's the answer.

MR. DE BRANTES: It doesn't matter whether it passes here or not.

MS. IBRAGIMOVA: The results are 3 votes for yes, 14 votes for no. This measure does not pass.

CO-CHAIR KOTTKE: Since lunch isn't until 12:30, let's move on to 2749. George, this is proportion of patients with arrhythmia that have potentially avoidable complications.

MEMBER RUGGIERO: George and I spoke about this last evening, and once again, we had a little difficulty wrapping our heads around it. I think this morning's discussion helped a lot.

To talk about the evidence, they do cite some papers here, about seven articles, which show that there is a correlation between the Type I PACs, and giving reference to patients who then subsequently have events related to that.

Then I looked at other chronic medical conditions and subsequent events related to having chronic medical conditions. I think that's the correlation that they're drawing. I think it does show what they're trying to look for going forward with the measure.

CO-CHAIR KOTTKE: Somehow, I got a disconnect in there. Can you restate? Do you think there's evidence?

MEMBER RUGGIERO: If I'm reading this correctly -- and once again, it was a little bit complicating. They do give a bunch of references here which essentially say why they're trying to go ahead and look at patients who have arrhythmias. The one that they state specifically is atrial fibrillation, if I'm reading this correctly.

In the patients who have atrial fibrillation, with the data that they show, they will have complications that occur related to the atrial fibrillation, which is the Type I PAC, which they describe. They also show that other chronic medical conditions will have similar PACs. So they draw the correlation that you will -- with A fib, you're expected to have PACs that are directly related to atrial fibrillation, and that with other chronic medical conditions, you will have PACs.

They're just showing a correlation that whatever your chronic medical condition is, you'll

have adverse events, some of which will be related to the medical condition, some of which will just be random. Is that sort of what I'm reading from it? That's what they show as far as evidence, just showing that there are papers to support that if you have an arrhythmia, or if you have a chronic medical condition, you're going to have adverse complications that will occur secondary to the arrhythmia, or just secondary to having a chronic medical condition. Is that correct?

CO-CHAIR KOTTKE: George?

MEMBER PHILIPPIDES: Agreed. We've already clarified the issue of facility level versus individual level. That was one question that I had. Again, it seemed to me that the PAC 1 group definitely related to physician performance. Many of the PAC 2 issues, but many not. For that reason, I have the same concern about what level we're really evaluating here.

CO-CHAIR KOTTKE: Further discussion? Hearing no further discussion, let's vote on the evidence for 2749, proportion of patients with an

arrhythmia that have potentially avoidable complication during the episode time window.

MS. IBRAGIMOVA: Importance to measure and report 1A evidence structure process intermediate outcome, 1 high, only eligible if QQC submitted, 2 moderate, 3 low, 4 insufficient.

CO-CHAIR KOTTKE: There's no QQC required. It's an outcome. You know how to vote.

MS. IBRAGIMOVA: So importance to measure and report 1A evidence, health outcome or PRO, 1 yes, 2 no.

(Voting.)

MS. IBRAGIMOVA: The results are 5 votes for yes, 12 votes for no. I believe it does not pass.

CO-CHAIR KOTTKE: Mary says I'm on a roll -- 2751, proportion of patients undergoing angioplasty procedure that have potentially avoidable complication. Joe?

MEMBER CLEVELAND: We've obviously had a robust discussion about a few things. I guess, at least taking this from the evidence perspective,

I think is a -- I echo some of the sentiments that fever maybe is not equally as weighted. That's one of the things maybe we can talk to, I'll focus on the evidence.

I think, again, from purely an evidence standpoint, yes, there is evidence that avoiding complications after defined intervention, like an angioplasty, is going to result in better outcomes. So I think from that standpoint, you could say there is evidence --the developer presents evidence in this, again, in a similar fashion, with some of the other trials, or some of the other Type I/Type II complications, etc.

But there are evidence from some of the commercial payers that they cite to suggest that avoiding these complications in other models, it seemed to follow logically that if you avoid a complication, it's going to be a better outcome. I think the evidence for this I support.

CO-CHAIR KOTTKE: I'm the other reviewer. I have nothing more to add. Anybody have anything they wish to say about -- yes, Leslie?

MEMBER CHO: I think this is a different measure than the ones we have been looking at. This is 90 days, which is -- it's a much more shorter time than the one year. I think that for me, this measure makes much more sense because it's a self-limited, time-limited thing related to a particular procedure. I don't want everybody to think that this is a me too measure on the other measures.

DR. RASTOGI: To that point, I want to add pacemaker was 30 days.

CO-CHAIR KOTTKE: Liz?

MEMBER DELONG: So how many -- I'm sorry, I don't have it in front of me. How many complications are included in this one?

DR. RASTOGI: It depends how you count it. If you look at the categories, sepsis is one, but it may have 180 codes. Then you will have 180 lines of complications listed.

MEMBER DELONG: So the stubbed toe is still in this one?

CO-CHAIR KOTTKE: No, it's never been

in there. Also, there's not really 800 --

MR. DE BRANTES: Stubbed toe isn't on this one.

CO-CHAIR KOTTKE: There's not really 800 complications because -- there may be 10 or 12 or 15 for ways to describe a particular event. We all know about coding, so --

DR. RASTOGI: And with ICD-10, it'll be 500 times more.

CO-CHAIR KOTTKE: Yes, including knitting injury -- hand injury by knitting or something. There's a lot, but it's under 100. Yes, Joel.

MEMBER SPANGLER: Tom, I had a question. This is listed as a composite and outcome whereas, the other ones were all just outcomes. Is that correct, the previous ones?

DR. RASTOGI: They're all composites.

MEMBER SPANGLER: Because I don't think -- okay.

(Simultaneous speaking.)

CO-CHAIR KOTTKE: -- any or none

composite, and so --- Joel?

MEMBER HOLLANDER: I was just going to say this also seems a little more reasonable to me to attribute to an individual clinician, someone who does the PCI. They have some responsibility for the patient for a short term afterwards. It's not a whole care team. At least from that point of view, I could think about this one as being slightly different than the other ones.

CO-CHAIR KOTTKE: Leslie?

MEMBER CHO: I'm sure it changes between this Type I and Type II PACs, your numbers for the thing. Can you just -- on this measure alone, can you give me the numbers for Type I and Type II PACs?

DR. RASTOGI: The PAC rate?

CO-CHAIR KOTTKE: I think it's a PAC count. Aren't you asking for the count, the number of Type I --

MEMBER CHO: No, I just want what's the median --

DR. RASTOGI: Between one and two?

Okay. I'll give you in a minute, so if you want to continue your discussion, but I'll pull it up.

To the extent it doesn't come in the data, don't worry about some of this. Because if it's 0 percent, it won't hurt any of you.

CO-CHAIR KOTTKE: So are we ready to vote on evidence --

DR. RASTOGI: Okay, Type I PAC for PCI was 37.5 percent, Type II was 21.6 percent.

CO-CHAIR KOTTKE: Kristi?

DR. RASTOGI: Overall, 47.5 percent, so there could be overlaps. The same patient may have Type I, as well as Type II, and providers need to focus on reducing all of them.

CO-CHAIR KOTTKE: Kristi, then George.

MEMBER MITCHELL: You might have already clarified this. Also looking at this at a population level, so even broader than the health system, but perhaps a state or national level because in the write-up it has that.

MR. DE BRANTES: Yes, we have looked at that. In fact, we're looking at county

distributions of rates of avoidable complications and the differences by county, which is quite informative. Our goal is to help consumers make decisions. County-level information doesn't help consumers make decisions.

With our colleagues in the employer community and the consumer advocacy groups, what we're looking to do is help consumers make decisions, and at the same time, provide useful feedback to providers on these rates of avoidable complications. Our experience to date -- which has far, far more evidence behind it than the measures you just approved for e-measures by the way -- shows that it's useful for consumers, and it's incredibly useful for providers, as well.

CO-CHAIR KOTTKE: George.

MEMBER PHILIPPIDES: Yes, comment on the questions about the number of PACs. I actually don't know that I care so much about the number. If it's 700 that are really good clinical indicators of care, then that's great. Maybe there should be 900, as long as they're all good.

Similarly, if one or two of them are a little bit wacky to us, like the stubbed toe, which is I know is not in there, but the other 800 are really good clinical indicators, that's not going to really change the fact that overall, the tool is still useful for discerning care from one system to another, or in this case, one provider to another.

I'm less interested in the specific number and an occasional weird one because I don't think that's going to change the fact that it's useful or not. What I care about, as you know, is what level it's at, which is my two cents.

CO-CHAIR KOTTKE: Tom James.

MEMBER JAMES: You can tell me, Tom, whether this comment belongs under validity or here, under evidence. The question I have has to do with what may be the unintended consequences of this, just as with the coronary artery bypass grafting reports that came out 15 years ago or so, where those physicians who had higher complication rates ended up leaving New York and Pennsylvania.

In part, they were saying it was because

they were dealing with patients that had greater -- who were more seriously ill, but you end up with the same kind of thing that perhaps we would see a total reduction in PCI done, and that could be good or bad, as a result of having a measure that looks at avoidable complications. Is that a validity or is that an evidence question?

CO-CHAIR KOTTKE: I don't know. Let me think about that with my hypoglycemia. Sana.

MEMBER AL-KHATIB: One thing that George said makes sense to me, in terms of it's not just about the number of these events or these complications, but it's also about the type of the complications.

I will be stressing this more when I talk about the pacemaker and ICD, but I find a lot of the ones that they listed for pacemaker and ICD not clinically -- not making sense clinically. For example, a lot of these seem to be pre-existing conditions. Why are we penalizing the physicians for something that these people had? Some of them are congenital issues, if I'm looking at the right

list. Some had to do with flushing, some with pallor, all these kind of things. If any other physicians decide to use those codes when they see a patient, maybe not even in your clinic, you'll end up dinging the physician. That's the concern that I have.

MR. DE BRANTES: Can I offer a response?

CO-CHAIR KOTTKE: Yes, Francois.

MR. DE BRANTES: This is also a response for you, Tom James. Hi, by the way.

The measure is a risk-adjusted rate. We're not dinging anyone. We're calculating comparative performance of providers on a risk-adjusted rate. To the extent that your rate is the same everyone else's, then how exactly are we dinging you?

I think you're confusing a process measure from an outcome measure. Here, we're calculating a risk-adjusted rate of potentially avoidable complications. As we showed, there's considerable variability in that rate. The extent to which someone's risk-adjusted rate is

significantly higher than someone else's, then our evidence suggests, and our data analyses suggest, there's something very different going on with the management of the patients with that particular provider than with other providers. We're not dinging anyone. We're providing comparative performance information on a risk-adjusted basis.

MEMBER JAMES: I appreciate that, and I recognize, too, that different risk-adjusting models end up with different results. I'm just thinking in terms, though, of reporting this kind of information, will that end up causing a net reduction in PCI because people at higher risk, even if there's risk stratification, physicians may be less likely to go ahead and perform this procedure. That could be good or bad. I don't have a value judgment on that.

CO-CHAIR KOTTKE: Let me go to Mike, then Leslie.

MEMBER CROUCH: You keep talking about the measure being useful. It occurs to me that may be possible because the wacky codes aren't ever

used. You've got several hundred codes that I don't think I've ever used in my life as 35 years a physician. I'm wondering why you haven't gotten rid of the ones that hardly ever appear or never appear. That may be cluttering our analysis of the measure.

DR. RASTOGI: It's the potential for gaming.

MR. DE BRANTES: This is the unending issue of both organizations of codes, as you know, which is a complex issue for anyone who's in the measure development community, as well as the -- even with ICD-10, the lack of specificity in many measures.

What we're trying to do, again, here is create a risk standardized rate of comparative performance. So to your point, if some codes are never used, it just never even comes into play in the calculation of that particular rate. But there's a reason why you have 180 codes, for example, that are grouped around sepsis because any one of those can be used at any point in time.

So there's an unintended consequence -- very clear unintended consequence of thinning things down to what you might consider to be purely acceptable, which is you substitute one code for another on the billing, and we don't want these games. None of us want these games. The goal is not to try to game a measure or not game -- it's really about how do we improve overall? How do we understand why there are these significant differences -- and there are very significant differences if you look at the data that we supplied -- and whether or not the measures are useful to providers in understanding, then how to tackle the variability that they see and that we see.

That's what these measures have been used, and that's what they -- that's where their power comes from. It's because it reveals the variability that exists, and it provides an opportunity, then, to understand the causes of that variability, and then to act on it. That's the reason why we're not going to thin down the codes

because we don't want people playing games on codes. We want, just like what you want, which is process improvement for care improvement.

DR. RASTOGI: Just one clarification. The pallor/flushing, it was in the typical list. It's not in the PAC list. I checked the workbooks. So that we gave both lists because the episode consists of typical and complications, and anything outside this is not part of the episode. So the whole ICD-9 book is so big, but only these are included.

CO-CHAIR KOTTKE: Yes, Liz, and then Leslie. Then I think it's time to vote.

MEMBER DELONG: I don't think any of us is not invested in promoting healthcare performance, but to endorse a measure, we want to make sure that we're dealing with quality performance measures. I'm not sure that endorsing a measure that has a variety of PACs that differ a great deal, from sepsis to something that's relatively trivial, is -- you can call it dinging, or you can call it risk standardizing, but you're

comparing people.

If you have a huge number of inconsequential PACs versus somebody with a few serious ones, you're going to ding or compare that provider who has a lot of trivial ones unfavorably to the one who has a few significant ones, and that bothers me.

MEMBER CHO: Final question. Why is PCI 90 days and pacemaker 30 days?

DR. RASTOGI: We worked with the clinical working groups, and that's what -- the number they picked. The pacemaker is basically more an outpatient type of procedure, and they wanted just to be held accountable for 30 days. We said fair enough.

PCI could have the same argument, and it could be 30 days, but we saw that sometimes after PCI, people were coming back for repeat PCIs, and they were delaying the repeat beyond 30 days. So wherever you make the cut, they'll just do the second one afterwards. We discussed with them about staging of a PCI, so quite a bit of discussion

was around that.

CO-CHAIR KOTTKE: Are we ready to vote? We are ready to vote on -- I guess it's evidence, we're still on 1A evidence.

(Simultaneous speaking.)

MS. IBRAGIMOVA: Importance to measure and report 1A evidence, health outcome, or PRO, 1 yes, 2 no.

CO-CHAIR KOTTKE: This is PCI.

(Voting.)

MS. IBRAGIMOVA: The results are 11 votes for yes, 6 votes for no, so this measure passes. We'll move on.

CO-CHAIR KOTTKE: Okay, Joe, opportunity for improvement.

MEMBER CLEVELAND: The performance gaps were calculated from PROMETHEUS, a large administrative claims database, period of study from April 2012 to December 17, 2014. Data was present for about half of the PCI -- 5,000 episodes, 5,898 of 10,000 PCI episodes in over 3 million beneficiaries.

Both unadjusted and risk-standardized PAC rates have a median of 50 percent, with an interquartile range of 44 percent to 55.6 percent. To me, that suggests there is a significant performance gap that exists.

CO-CHAIR KOTTKE: I would agree.

MEMBER CLEVELAND: No data was presented for disparities.

CO-CHAIR KOTTKE: Linda.

MEMBER BRIGGS: I guess this goes back to what Liz was saying about how's this data reported? I would agree that there is an opportunity for improvement here. It's like when patients -- when you tell somebody that their ejection fraction is 55 percent and they think it's out of 100, then there's their concern.

If you tell -- if you report a rate of 50 percent unadjusted complication rate on a provider or a facility, if they're not looking at that as comparison to everybody else, then the consumer -- if this is meant for consumers to look at -- is going to think that's really a bad score.

It might not necessarily be a bad score. Is that right?

CO-CHAIR KOTTKE: There is variation around --

PARTICIPANT: Yes.

CO-CHAIR KOTTKE: There's variability, so I would interpret that as room for improvement. Other comments, or can we vote? Let's vote on performance.

MS. IBRAGIMOVA: Importance to measure and report 1B performance gap, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. IBRAGIMOVA: The results are 7 votes high, 6 votes moderate, 2 votes low, 2 votes insufficient.

CO-CHAIR KOTTKE: Specifications, reliability, reliability testing, Joe.

(Simultaneous speaking.)

CO-CHAIR KOTTKE: I'm sorry. I jumped the rails here. Quality construct.

MEMBER CLEVELAND: We've had a very

robust discussion about the quality construct. In this quality construct, too, each PAC is equally weighted, so again, I won't use stubbed toe, but it does pose procedural fever, which could -- I understand the developer in wanting to cast a broad net, but that's equally weighted with hemopericardium and other things that could be potentially viewed as much more significant.

I think that's something, again -- I don't know if we want to have more robust discussion about that or not, but to throw that out there.

CO-CHAIR KOTTKE: Taking the pragmatic side of how would we create a weighting system that is other than arbitrary, recognizing that a post-op fever is not the same as a hemopericardium that -- or a hemothorax that requires a chest tube.

MR. DE BRANTES: Yes, so you can imagine we had a tremendous amount of discussions about the weighting. We agree that whatever weighting schema ends up by being, of course, arbitrary, and then subject to anyone's arbitration.

There is a potential proxy for severity

or weighting, if you will, that we considered, which is looking at the relative cost. The challenge there is that while you can capture those for ambulatory-based avoidable complications, you cannot capture those for complications that occur during a hospital stay because all of those costs, as you know, are lumped into one single bill, and you can't differentiate the cost of the individual components.

So the more we looked at the different ways of potentially weighting them, it came down to there's virtually no way of doing it in a manner that could be deemed remotely objective and so, therefore, no weighting seemed like a better option than weighting.

MEMBER DELONG: I agree that any kind of weighting would be arbitrary and not valid, but then again, you're lumping fever with sepsis, and is that valid?

PARTICIPANT: It seems fever may be an indication of sepsis. At least, that's what we learned in medical school.

CO-CHAIR KOTTKE: Other discussion around quality construct? Let's vote on quality construct.

MS. IBRAGIMOVA: Importance to measure and report 1C composite, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. IBRAGIMOVA: The results are 0 votes for high, 11 votes for moderate, 2 votes for low, 4 votes for insufficient.

CO-CHAIR KOTTKE: Reliability specifications, reliability testing.

MEMBER CLEVELAND: So the developer tested reliability of the performance measure score. Analysis was with a data binomial model, and also signal-to-noise analysis. I believe -- and correct me if I'm wrong -- analysis was carried out only for the facility for this measure. Is that correct? Which then encompasses, I think, something that we're a little more comfortable with.

The one thing to note, though, is that

of the 565 facilities initially included in the dataset, to the points earlier illustrated, because of the spread of things, it left 41 facilities for analysis. So it gets down to kind of a volume relationship of the more PCIs a facility did, the better the reliability was, with a 0.74. Whereas, for the ones that you put the cut at just ten PCIs for a facility, their reliability was not great. It was 0.5. I think we've heard that expressed.

I guess one question is how will this reliability exist across all facilities? I think that's just something that is, unfortunately, not able to be answered, other than just the fact that there's signal-to-noise in this.

CO-CHAIR KOTTKE: This is really a bit of a conundrum that the low-volume places that you can't really assess, I would suspect, may be the ones with the biggest problems. You must have thought about that.

MR. DE BRANTES: Yes, we did. I think what we have all suffered from are measures that patently fail reliability testing and that are

still widely used, sometimes endorsed, and around which inferences are made about the performance of physicians or facilities in a completely inadequate way.

So our approach to this issue, as I stated before, is that these scores should only be calculated on the providers for whom the -- where the dataset reveals reliability. We suspect that the answer would be different, by the way, again, in a Medicare dataset because you have a lot more volume distributed amongst a lot of facilities.

For commercial populations, you're right, we don't know. So is it better to have a less reliable measure that creates a poor inference, or in this case, you probably would end up with -- we looked at it. You would end up with undifferentiated performance, which we think is actually a worse signal to provide than no score at all.

Then it's up to the, I think, consumers, the payers, to draw inferences about do you actually want to go to a facility on which a score -- a

reliable score of performance cannot be calculated because they have inadequate volume than going to a facility for which a reliable score can be calculated?

CO-CHAIR KOTTKE: I have another question. I know you don't want to go here, but does the narrowing the scope of codes change reliability by -- doesn't do anything? Okay. Does Linda have a question?

MR. DE BRANTES: To expand on it, because as we have all discussed, it's the stuff that matters that ends up by popping to the surface, and the stuff that doesn't matter doesn't pop to the surface and ends up by being irrelevant to the -- but as we also discussed, thinning down the code set creates a potential unintended consequence that we're going to try to avoid.

DR. RASTOGI: Just one more point there. The PAC drill-down reports, you can see the important stuff floats to the top.

CO-CHAIR KOTTKE: Then I had one other question. In your document, you say that anybody

can use this if they can -- but is there some mechanism to -- is this self-protective, that trying to use it in small groups or underpowered situations that you're unlikely to turn up anything anyway? The question is somebody applies this in ten doctors, and in one doctor, they sort of say, this doc's high, kick him out. Kick out the highest doc.

DR. RASTOGI: Our recommendation is to do the reliability testing first, and if the reliability is more than 0.7, only then the measure should be used. If not, we just give volume information, and the consumer can easily see.

MR. DE BRANTES: As we all have experienced over the past years of, again, endorsed measures being used for whatever purpose without any constraints around them, people will do whatever they want with them, sometimes perfectly inadequately. Again, we're very clear in our application that these measures should be used after having conducted a reliability test in the measure set that you're using to assess

performance.

CO-CHAIR KOTTKE: Right. Yes, people have used decision rules inappropriately for centuries. Okay, are we ready to vote on reliability?

MS. IBRAGIMOVA: Scientific acceptability of measure properties, 2A reliability, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. IBRAGIMOVA: The results are 0 votes for high, 11 votes for moderate, 4 votes for low, 2 votes for insufficient. This measure does pass reliability.

CO-CHAIR KOTTKE: Validity, Joe.

MEMBER CLEVELAND: The measure uses statistical risk model, again, with 170 risk factors imputed. Developer conducted a very thorough systematic assessment of face validity using multi-specialty working groups, focus groups, and comparisons to other national accountability measures.

There were no empiric results provided for the face validity tests, but I think the only threats to validity were, again, some of the things we talked about already, the exclusions with, again, some of the facilities falling out of the reliability tests.

The performance of the risk model was determined with a split sample method, however, too, by estimating model coefficients using a developmental dataset and apply these coefficients, so C statistics were good, 0.803 and 0.792. Anyway, I think from the standpoint of validity, there is validity in these data. I don't have a problem with it.

CO-CHAIR KOTTKE: I would agree. Anybody want to comment otherwise? SDS?

MEMBER CLEVELAND: Not done -- so.

CO-CHAIR KOTTKE: Anybody care to talk more about SDS? Apparently not.

MS. WILBON: It might be useful, too, to have the developers -- I know you guys mentioned a little bit about the physician at Montefiore that

did some -- oh, I'm sorry -- did some study with the measures at Montefiore.

So maybe you guys could just talk a little bit about your thoughts about it and give the committee an idea about what you may or may not have considered, in terms of adjusting for SDS factors.

DR. RASTOGI: As Spencer mentioned earlier, we don't have that data right now to link to administrative claims data. Once it's available, it would be nice to look at it both ways.

CO-CHAIR KOTTKE: So really, it's not possible.

MR. DE BRANTES: You can look at, and we have looked at, for example, the usefulness of zip codes as an indicator of sociodemographic status. As soon as you get into high density areas, even zip plus four is completely useless as an SDS predictor. I think there's a fair amount of studies that show that.

If you go back and you look at the work that folks in Cleveland have done around the SDS adjustments for diabetes performance measures,

what they've found is that the insurance status ends up by being the most strongly correlated element or variable with sociodemographic status. Because here, we're looking at commercially-insured populations, that kind of ruled out comparing Medicaid, say, from commercial.

So there isn't, and then there's the, I think, overriding philosophical issue, which I think Larry Casalino encapsulated very well years ago, which is to the extent there is a -- and I think we know that in certain populations there are differences in results -- that the goal of SDS adjustment should not be to adjust away the differences, but rather to create a baseline differential and to work hard to squeeze that differential out.

At some level, I think we share that ambivalence about -- beyond just the fact that it's not necessarily doable in commercially-insured populations, we also share the ambivalence that adjusting for those differences by quote, unquote, eliminating them, doesn't serve the population that

is on the receiving end of that very well.

DR. RASTOGI: Just to add to what Francois said, as you know, our models are in use for payment purposes. On the payment side, we do want to give an additional allowance to doctors who care for the low SDS patients.

CO-CHAIR KOTTKE: Okay, ready to vote on validity?

MS. IBRAGIMOVA: Scientific acceptability of measure properties, 2B validity, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. IBRAGIMOVA: The results are 0 votes for high, 12 votes for moderate, 3 votes for low, 2 votes for insufficient.

CO-CHAIR KOTTKE: Okay, showing that I can learn, I will not forget empirical analyses to support the composite. Joe.

MEMBER CLEVELAND: I think we've had a fairly robust discussion about this. I don't know if I have anything else to add.

CO-CHAIR KOTTKE: Are they there? Are

the analyses there?

MEMBER CLEVELAND: I think there are some analyses there to support it. Again, I think that, again, the equal weighting in the composite is something that is -- you just have to take on faith that either you weight or you don't weight. If we weighted it, I agree that it would be completely arbitrary.

CO-CHAIR KOTTKE: Am I correct that the only thing -- they have to present the distributions, and that's it, right?

MEMBER CLEVELAND: Right.

CO-CHAIR KOTTKE: But that's a minimum standard, which they've done.

MEMBER CLEVELAND: They've done.

CO-CHAIR KOTTKE: Okay, we have to vote on that, and they've satisfied that criteria.

MS. IBRAGIMOVA: Scientific acceptability of measure properties, 2D composite, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. IBRAGIMOVA: We had another

technical difficulty, if we could all just recast our votes.

(Voting.)

PARTICIPANT: Hand vote.

CO-CHAIR KOTTKE: Hand vote time. Empirical analysis, 1 is high, how many people high?

PARTICIPANT: Are we going to hand vote?

CO-CHAIR KOTTKE: We're going to have to hand vote on this one. We're having technical problems. Seeing no highs, moderate?

MS. MARINELARENA: 11 moderate.

CO-CHAIR KOTTKE: Low -- 3 lows.

MS. MARINELARENA: 3 lows.

CO-CHAIR KOTTKE: Inadequate.

MS. MARINELARENA: Insufficient?

CO-CHAIR KOTTKE: Insufficient, sorry, 2.

MS. MARINELARENA: 2 insufficient. Does that pass, Laura?

MS. IBRAGIMOVA: Yes.

CO-CHAIR KOTTKE: Yes.

MS. MARINELARENA: Okay, it passes.

CO-CHAIR KOTTKE: Feasibility.

MEMBER CLEVELAND: If the data sources are readily available in electronic sources to large administrative claims database, there's access, so I think it's feasible.

CO-CHAIR KOTTKE: Agree. Anybody have any comments? Seeing no comments, let's -- have we solved our -- okay, feasibility. Let's vote on feasibility.

MS. IBRAGIMOVA: Feasibility, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. IBRAGIMOVA: The votes are 8 for high, 7 for moderate, 1 for low, 1 for insufficient.

MR. DE BRANTES: Just a technical question. If the criteria are data generated during -- so the data are readily available, and they're in an electronic source, and it can be easily implemented, how can it be insufficient?

CO-CHAIR KOTTKE: Does anybody care to fess up? So that will remain a rhetorical,

unanswered question.

MR. DE BRANTES: Thank you.

CO-CHAIR KOTTKE: Joe, usability and use.

MEMBER CLEVELAND: So this is a claims measure that's used now in programs for payers states' external quality reporting, so it's already in use. There are no, I'd argue, unidentifiable, unintended consequences should this -- we've already had at the measure at clinician/group level. I put that as the only questions, but I think it's usable.

CO-CHAIR KOTTKE: I would agree. Time to vote? Any other discussion? Seeing no other discussion, let's vote on usability and use.

MS. IBRAGIMOVA: Usability and use, 1 high, 2 moderate, 3 low, 4 insufficient information.

(Voting.)

MS. IBRAGIMOVA: The results are 7 votes for high, 7 votes for moderate, 2 votes for low, 1 vote for insufficient information.

CO-CHAIR KOTTKE: We'll vote on overall endorsement or recommendation is suitable for endorsement. This is a PCI. There seems to be a fairly strong comfort with the measure. Let's go ahead and vote.

MS. IBRAGIMOVA: Overall suitability for endorsement. Does the measure meet NQF criteria for endorsement, 1 yes, 2 no.

(Voting.)

MS. IBRAGIMOVA: The results are 10 votes for yes, 7 votes for no. That's gray zone.

CO-CHAIR KOTTKE: Leslie, are we going to have lunch, or are we going to finish this?

MEMBER CHO: We won't torture you any longer. We will break for lunch now, and we'll come back at 1:30, and we'll continue on with the rest of the HCI3 measures.

MEMBER SPANGLER: Tom? Sorry, what happens with the gray zone?

MS. WILBON: With the gray zone measure, all of the measures that you guys vote on today go out for comments. So this measure will

come back to the committee at the post-comment call, and you'll have the opportunity at that time to consider the comments, and you'll be given the opportunity to re-vote again before it goes to the CSAC.

(Whereupon, the above-entitled meeting went off the record at 12:56 p.m. and went back on the record at 1:31 p.m.)

CO-CHAIR GEORGE: Okay, we're going to go ahead and get started with the next measure, 2752, Sana and George.

MEMBER AL-KHATIB: This measure has to do with proportion of patients undergoing pacemaker defibrillator implantation that have a potentially avoidable complication. What I would like to do, since we've had robust discussions about similar measures, is to ask the developer a couple of questions.

First, I do want to acknowledge all the hard work that went into developing this measure. I'm sure this was not easy, especially having worked on all these other measures, so I do want to

acknowledge that. But in terms of looking at the PACs -- by the way, for an electrophysiologist, I see PACs, I'm thinking premature atrial contractions, but anyway.

The Type I PACs I fully understand, and I fully accept. Where I think I still struggle is with the Type II PACs that you submitted. I just want to make sure that I am reading the Excel spreadsheet correctly, that you included 868 PACs. Am I reading that correctly?

DR. RASTOGI: Yes. As I mentioned earlier, these are individual codes. If you look at that Column E, that gives the PAC name, and then the codes that relate to that, but yes.

MEMBER AL-KHATIB: So I certainly accept the fact that if there are certain conditions that I can remotely relate to the procedure that I think that's certainly justifiable, certainly makes sense to me. But when we're looking at complications that have no correlation whatsoever with the procedure that was done, that's where I start to struggle with these performance measures.

So in looking at the list here, there are so many things listed, like poisoning. You listed close to 217 forms of poisoning that have nothing to do with the procedure. A lot of these concerns have been voiced before, so I don't want to belabor the point, just to say that remains to me the main concern that had these conditions been more relevant, more related to the procedure itself, I, personally, as a physician, would be much more accepting of these performance measures.

Then the other point that I want to clarify is in just looking at the description, the 30 day versus the 90 day. I know Leslie asked about that earlier. You mentioned that the clinical experts in your group favored 30 days for this. As somebody -- I am an electrophysiologist, and I actually devote close to 90 percent of my clinical practice to devices -- pacemakers, ICDs. I love doing them.

I totally agree with you that we need to have performance measures in place that can help improve the outcomes of our patients. I'm sure we

all agree that we want better outcomes for our patients, but I also want to be cognizant that we're doing it right. So in terms of looking at the 30 day versus 90 day -- in fact, our group has published on this quite a bit.

We found that a 90-day timeline makes more sense because you're capturing all kinds of complications, not just those early complications. A lot of infections don't declare themselves until you're past that 30 days. So I would worry that not only are we capturing things that we may not want to capture, but now we're also missing out on important outcomes that may not manifest themselves until 90 days or later.

I would remind the group that when, yesterday, we reviewed the performance measure from ACC on complications, you may recall that the complications -- some were within 30 days. Some of the measures were within 90 days. They had a technical expert panel that informed them on that issue. I think there might be some consensus, I would say, at least in the EP community, that 90 day

might make more sense. So your thoughts on that, please?

DR. RASTOGI: Going back to your first question about poisoning, if you look at the PAC drilldown reports, one of the tabs -- it's after the PAC overview tab -- there are six patients out of all these who had the pacemaker who had that poisoning stuff. The fact is that many of these bad things happen in a facility setting. That's why they are Type II PACs.

Type I is where the provider is directly responsible, while Type II is the system. The nurse gave the wrong injection or something like that. I personally have seen, even at Mayo Clinic, something like that happen. So I don't want to say that we want to limit the number, but then obviously, when you're comparing provider performance, systematically, if somebody is seeing many more poisonings than the other, then there's something to say about it, but otherwise, it'll be like in that baseline that somebody was saying, that it won't surface up, it'll just be that little

variability across.

I can hear that you may not have anything to do with if the nurse gave the wrong medicine when you wrote the right one, but it's a matter -- and the pacemaker currently is being tested at the facility level. So we did want to have all the potentially avoidable complications pulled together and not limit that. Then your second question was --

MEMBER AL-KHATIB: 30 day versus --

DR. RASTOGI: Yes, 30 days, yes. You're right. We have gone back and forth in that. When we did, say, total knee replacement episode, initially it was six months, then we made it three months, and even, as you know, in the CMS, both options exist, whether people want to do a 30-day accountability or longer periods of accountability.

So all that, the debate is still -- the jury's still out. There's no clear-cut answer. Yes, some delayed infections would show up much later. We can certainly broaden it. How much

noise will we be pulling? How much signal will we be pulling? It's a trade-off. At the time when the measure was created with the clinical working groups, the feedback we got is let's limit it to 30 days. If you do 90, then again, the question is why should we hold it? The discussion goes both ways.

MEMBER AL-KHATIB: Thank you for your clarification. I remain concerned regarding the comprehensiveness of the lesson that they provide. Again, a lot of those things, I don't see them at all relevant to this encounter.

A lot of the poisoning things that you mentioned had nothing even to do with medications that we would use around the placement of the device or in any way related to the procedure itself. That's why when we try to talk about evidence and try to talk about this association, I just find that the evidence is lacking.

CO-CHAIR GEORGE: George, do you have any -- is there anything in the way of evidence to support the 30 day or 90 day?

MEMBER AL-KHATIB: Yes, up to 90 days,

definitely. As I said, infections, a lot of the infections, you're not going to capture them within the first 30 days. That is a concern of mine. A lot of these infections are not going to come back to your attention until past those 30 days, and you're not going to be able to capture them.

MR. DE BRANTES: We also do a fair amount of empirical testing in our datasets. Again, I think there are differences between Medicare patients and commercially-insured patients. As we looked at what happened to patients who underwent PCIs 30 days, 60 days, 90 days out, the links between the avoidable complications in that period of time were pretty strong; whereas, less strong for patients who had implantable defibrillators once you got past 30 days.

So there are other things that can create more noise in the data and that, therefore, dilute the potential impact, at least for that particular procedure in the datasets that we looked at. So again, this is dataset by dataset. We're

not talking about Medicare patients. We're focusing on commercially-insured.

MEMBER AL-KHATIB: What I would add to that is actually, there are certain billing codes that are specific to device infections. If you're just using a generic bacteremia or something like that, I agree that the whole point about introducing noise makes sense to me, but there are specific codes, even with the ICD-9 codes -- specific codes to device-related infections, pacemakers, ICDs, what have you.

CO-CHAIR GEORGE: Any other comments on the evidence? If not, we'll move to a vote on the evidence.

MS. IBRAGIMOVA: Importance to measure and report 1A evidence, health outcome or PRO, 1 yes, 2 no.

(Voting.)

MS. IBRAGIMOVA: Tom, can you please cast your vote via chat or text? Operator, is Tom James still on the line?

OPERATOR: His line is still connected.

MS. IBRAGIMOVA: Thank you. The results are 9 votes for yes, 6 votes for no. I believe that is gray zone.

CO-CHAIR GEORGE: We'll continue with the opportunity for improvement.

MEMBER AL-KHATIB: The developer presented data regarding -- I think they used the PROMETHEUS administrative claims data between 2012 and 2014 and talked about these episodes and showed that the unadjusted rates, the median was 46.8 percent. There was a range, as well.

When they risk standardized these rates, the numbers were not -- didn't change much. Clearly, there is variability. The number is pretty high. But I worry that even after risk adjustment, that rate didn't change much, so it leads me to question -- and maybe this is a point that we will have to tackle later. It leads me to question how well the risk adjustment is working.

CO-CHAIR GEORGE: Other comments on the opportunity for improvement?

MEMBER AL-KHATIB: They didn't provide

any information on disparities. I forgot to mention that.

CO-CHAIR GEORGE: Thank you. All right, we'll vote.

MS. IBRAGIMOVA: Importance to measure and report 1B performance gap, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. IBRAGIMOVA: The results are 2 votes for high, 11 votes for moderate, 0 votes for low, and 2 votes for insufficient.

CO-CHAIR GEORGE: We'll move on to the quality construct.

MEMBER AL-KHATIB: I think I've already talked a lot about the construct, so I'll just open it up to other people to voice their opinion.

CO-CHAIR GEORGE: Any comments from the rest of the committee? If not, we'll vote on the quality construct.

MS. IBRAGIMOVA: Importance to measure and report 1C composite, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. IBRAGIMOVA: The results are 0 votes for high, 8 votes for moderate, 5 votes for low, 3 votes for insufficient, and it's gray zone.

CO-CHAIR GEORGE: So we'll move on to specifications, reliability testing.

MEMBER AL-KHATIB: So starting with the specifications, we have talked about the 30-day window, so I'm not going to belabor that point.

I have a couple of questions for you about who that actually involves. Are these just new implants of pacemakers and ICDs, or are you also including replacements of devices, and what about patients who are getting cardiac resynchronization therapy devices?

DR. RASTOGI: Yes. The trigger list, the first tab, lists all the ones which are included, and they include all comers, so replacement, as well, and then the cardiac resynchronization devices also. I double-checked after your comment.

MEMBER AL-KHATIB: Thank you very much.

As I said, my concern is the PACs, how they were defined, the Type II one specifically. The Type I, I have no issues with at all, and the time frame.

But in terms of the testing -- I would actually remind people that the level for this one is also the clinician individual, but it can also be clinician group practice, clinician team facility integrated delivery system, so certainly at the level of the clinician there.

Reliability. What's mentioned here is regarding reliability testing, the measure is specified -- apparently what's required is that the measure is specified for use with individual clinician group practice team facility and integrated delivery system levels of analyses, though testing is provided just for facilities. It doesn't look like the developers were able to meet the expectation, although maybe I'm not reading that well. They just provided the testing data for the facility level. How is that viewed by NQF?

MS. WILBON: Generally, I think we're going to have to circle back with the developers.

We noticed a few discrepancies in the form earlier I think we're trying to still resolve. But generally, the policy is that the measure's endorsed at the levels at which it has been tested. For now, if the testing is only provided at the facility level, it would be endorsed -- recommended for endorsement --

MEMBER AL-KHATIB: If it gets endorsed.

MS. WILBON: -- for -- at the facility level.

MEMBER AL-KHATIB: Thank you. Then they talk about, in the reliability testing, that the beta binomial failed to produce statistically significant parameters, so they were, therefore, unable to calculate facility reliability scores. They were unable to report reliability scores, suggesting that statistically, the measure may not adequately differentiate between facilities in the current database tested. That's, I think, what was alluded to at the beginning, with the introduction.

So I think they ended up using a different database, with much smaller sample sizes,

and finding a reliability of greater than 0.7. Though the specific results are not provided, I know that I saw the number of greater than 0.7, but I don't know that we've seen enough information to know how that was derived.

DR. RASTOGI: That same reliability tab, if you slide down, you will see the ad hoc analysis details. If you have that workbook pulled up, you can slide it down. The first set is the standard dataset, and then the ad hoc analysis numbers are below.

You're right. The second sample, they just test -- we have so many datasets on which we are testing all this stuff. We used one huge sample for all the other measures, but for this one, it was a much smaller -- the first one had 3.2 million covered lives, while this one had maybe less than 2 million. For this, we had only 280 episodes that met the -- after provider attribution, the count of ten. Even in those 14 providers, the Alpha and Beta values were good. The provider variability was high. It ranged from -- it's 15 percent to 69, or

even 80 percent PAC rate.

That means in some people, it was very low, and in some, it was very, very high. So across provider variability was so high that the Beta binomial gave the between variance as 0.02. Because of that, the reliability for seven facilities was more than 0.7. For that, the sample size was at least 22 and above. So that's what we were saying. It varies from dataset to dataset.

In some datasets, you're seeing there's very little variability across providers, so the Beta binomial did not meet that criteria. But when we tested another sample dataset, the variability across facilities was huge. Now, these are regional datasets, so in Northeast, in one region, they may be very good performance. The variability may not be so good. In another dataset, you can see there's a lot more variability.

MEMBER AL-KHATIB: I don't want to put anybody on the spot, but I would like to hear your thoughts, Liz, if that's possible, about how they used a bigger sample size. I understand the

concept that it is dataset dependent, but could you actually show reliability with just 22 patients? Is that even -- I would think that's such a small sample size.

DR. RASTOGI: But the reliability of minimal sample size of 22, like in CAD, we were showing even at ten, a sample size of ten, we were seeing good reliability scores --

MEMBER AL-KHATIB: Across sites? Across facilities?

PARTICIPANT: Yes.

MEMBER AL-KHATIB: Okay, got it. Thank you. I don't have any further comments about the reliability. I still have some hesitation about the methodology.

MEMBER PHILIPPIDES: Sorry to keep asking this question. I just want to be sure. When we're voting on this, and this metric, in general, it's at the facility level, not the individual clinician level?

MS. WILBON: Yes, we'll be working with them to make sure the forms reflect that, but for

now, as long as -- what you see in front of you, in terms of testing at the facility level, is all that we can make recommendations on right now. It would be the facility level.

CO-CHAIR GEORGE: Does that apply just to the reliability or to everything?

MS. WILBON: It applies to the whole measure.

CO-CHAIR GEORGE: Thank you.

MEMBER AL-KHATIB: We were told yesterday, when we talked about a different measure, you have to do it based on what's in front of us -- the paperwork that's in front of us. They clearly say clinician level, individual level, based on the paperwork.

MS. WILBON: If I'm thinking about what you're talking about, that was a checkbox, but the testing -- what we try to do ahead of time is make sure that the boxes that they check, in terms of what they've tested and what they actually provide, in terms of testing results, align. In this case, if I understand correctly, they've actually provided

data and numbers showing they tested at the facility level, but checked more boxes than just the facility level. Is that correct?

MEMBER AL-KHATIB: Let's ask the developers. Do you want us to be voting on this at the individual level or at the facility level? Because when we asked this earlier, we were told that no, you want us to review it based on the individual level, so is that a change?

DR. RASTOGI: This particular measure -- this one and the PCI measure, they were tested only at the facility level.

MEMBER AL-KHATIB: Okay, thank you.

CO-CHAIR GEORGE: Any other comments on the reliability? All right, we'll vote.

MS. IBRAGIMOVA: Scientific acceptability of measure properties, 2A reliability, 1 high, 2 moderate, 3 low, 4 insufficient.

(Voting.)

MS. IBRAGIMOVA: The results are 0 votes for high, 9 votes for moderate, 5 votes for

low, 2 votes for insufficient. This is in the gray zone.

CO-CHAIR GEORGE: Validity.

MEMBER AL-KHATIB: I do have concerns about validity, as well, because there were no empiric results that were provided for the face validity, based on face validity tests. My understanding, that's actually required. Then even the developers --

PARTICIPANT: There's no requirement.

MEMBER AL-KHATIB: No requirement? Okay.

MS. WILBON: That's the minimum threshold. It's not required, but it's the minimum threshold.

MEMBER AL-KHATIB: So it's not provided. Then threats to validity. The developers actually did a good job providing information there. They said that -- they described patient demographic enrollment information and claims-based exclusions for the measure, and they said that nearly half of the

original population of patients was removed from the denominator because of exclusions that they applied. Then a significant number of episodes were eliminated from the measure due to exclusionary criteria, and then they provide the numbers for that. In terms of risk adjustment, they certainly did that, and that's certainly required, in my mind, at least, as a scientist. They needed to risk adjust for it. As I mentioned earlier, I am concerned about the practicality of adjusting any risk model for 170 risk factors. I'd like to know how those risk factors were actually chosen, and how do you see this playing out in practice? Is it practical to expect people to be adjusting for 170 risk factors?

DR. RASTOGI: Thanks for those comments. The first thing about exclusions, we did not have any exclusions. The piece that we have mentioned, and maybe it's a terminology issue that may have caused some confusion, those were selection criteria.

If a patient did not have claims for the

entire episode time window, so for, say, 30 days' worth of claims they had enrollment gaps, because it's a commercial dataset, in Medicare, you don't really see those gaps, but if the patient drops out of enrollment, then that gets kicked out. So we had selection criteria that were defined. Age has to be 18 plus. They have to match the enrollment, etc. The sample size that's selected -- the episodes that are selected met all those selection criteria. It's not really exclusion, but the other patients do not qualify because we don't have -- incomplete data, so to say. Then coming back to your risk adjustment question, the risk variables are collected in an historical fashion.

So again, through claims data, these risk factors are available or could be available, depending on how much dataset is available. So we have at least six months' worth of data that we require before the episode trigger to collect these risk factors. Yes, there are many, many risk variables, but as you can see in the risk model, not all of them have enough volume, and to the extent

they do, then they generate coefficients.

Now, there are many methods of doing risk adjustment. The approach that we applied was each risk factor, if it contributes towards the risk model, then it is kept in. But otherwise, it's a standard logistical regression model, so I'm not sure whether anybody has to collect those risk factors. It is just through the claim submission form these risk factors just get automatically generated through that.

MR. DE BRANTES: And another important point is that these risk factors are calculated dataset by dataset, for the same reason that you have differences in results on reliability testing dataset by dataset.

The epidemiology of patients changes and, therefore, the relative strength of any one of these risk factor variables is going to be very different from one dataset to dataset. So again, you want to be comprehensive in the number of risk factors that you look at, knowing full well that many of them will not have any impact on the severity

model, but some will. Those that do change dataset by dataset, so you can't -- you don't want to pre-judge which ones will have significance.

The other reason why that's important is because we don't think that imputing a single value on a regression model variable from, say, a normative dataset, however which way you describe it, does a good job at explaining differences in patient severity for a specific population studied. So all of these models are calculated dataset by dataset.

MEMBER AL-KHATIB: Thanks for the clarification. You know, I still see that this is such an extensive list.

Clinically I can make an argument for many of these variables that I don't know that, you know, you would want to adjust for.

And then as I was talking about, the performance gap data that you presented when you risk adjusted, those numbers didn't change much, which leads me to question the effectiveness of this risk adjustment.

And, you know, again we really need to try to get it right, because there may be some unintended consequences, because physicians when they see these kind of, you know, data, they may start, you know, cherry-picking the healthiest patients. And then patients who are sick may end up, you know, paying the price because they're not undergoing procedures that they need.

DR. RASTOGI: Yeah, and for the very reason we didn't want to restrict it to just a handful of risk factors, because then the cherry-picking becomes very important.

When there are so many risk factors that could be there, then every patient would have something or the other.

And the fact that the outputs did not change much, then you're right that maybe these risk factors did not have anything to do with the performance of the physician. The pacemaker is pacemaker and the outcomes may have been just not related to these risk factors.

So, the fact that the model didn't

change doesn't have anything to do with the presence of risk factors, but we did do the sample, the test and the validation data sets and we showed the statistics numbers and the predictive capability of these models. And the predictive power was very good.

CO-CHAIR GEORGE: Liz?

MEMBER DELONG: I wonder if you've done any validation of any of these models. When you run a model with 174 covariates, you do run the risk of a lot of overfitting.

And then you're going to apply that model to risk adjust and make assessments on hospital -- facility or physician performance.

DR. RASTOGI: So, those results have been provided in the section for validity testing.

And we have also shown the decile -- you know, breaking the data set -- the outputs into deciles and looked --- compared the observed from the expected. So, all those numbers are provided and, actually, we were surprised it performed so well.

I think it was the first decile that it did not, right? And then, balanced it. So, the c statistics for the test and the validation was 68.7 for the chi square.

CO-CHAIR GEORGE: Any other comments on validity?

MS. MARINELARENA: Before we move on, Amita, can you talk about did you look at SDS factors for this measure as well?

DR. RASTOGI: Yes. So, the same thing applies for the others that we didn't have availability of the SDS factors.

CO-CHAIR GEORGE: Let's go ahead and vote on validity.

MS. IBRAGIMOVA: Scientific acceptability of measure properties; 2b, validity. One, high. Two, moderate. Three, low. Four, insufficient.

(Voting.)

MS. IBRAGIMOVA: So the results are zero votes for high. Five votes for moderate. Five votes for low. Four votes for insufficient.

It does not pass validity.

This measure would technically fail. Do we want to continue?

MS. WILBON: I'd say at this point, let's go ahead and just continue on to the next measure.

The measure will go out for comment, like you said, and we'll bring back comments and continue the discussion if needed. Thanks.

DR. RASTOGI: Okay. One question. It's the same measure like the previous. I don't know why the validity counts as a different --- it's exactly the same.

MS. WILBON: So, it might be helpful to have some of the committee members maybe talk about what was different for this measure considering the other one passed, the PCI measure that was just before this.

MEMBER AL-KHATIB: I actually was consistent, because I voted exactly the same for both.

(Discussion off the record.)

MS. WILBON: Anyone care to share? Okay.

CO-CHAIR GEORGE: It might have to do with a different mix of people here in the room. We've lost some members, so.

MS. WILBON: Okay. Thank you.

MS. SPEAKER: It doesn't seem like that would be enough. I mean, it's dramatically different. Doesn't seem like a few would have made it ---

DR. RASTOGI: Yeah, like the insufficient and the low went up, right?

MS. SPEAKER: Yes, significantly. Right.

DR. RASTOGI: And it's exactly the same, so --

DR. BURSTIN: Right. And that would be fine if fewer people were supporting it moderate or high. What we're seeing is a shift to low and insufficient when, in fact, it's identical to the prior measure which you passed.

So, just, again, we get lots of concerns

for folks about inconsistency and it doesn't appear consistent.

MEMBER AL-KHATIB: From my perspective, actually, what I would say, probably, is that not enough information was included. And I'm not being critical at all. I'm just stating a fact. You know, with the initial measure on PCI, I felt like some of this information was not presented. I don't know how many factors. I wasn't the primary presenter for that.

So, I don't know how many factors that you guys looked at in the risk adjustment model, if they felt that those factors were clinically relevant or not.

So, and I don't know that they pointed out that even after risk adjustment the number didn't --- numbers didn't change. So it made me question the effectiveness of the risk adjustment.

I think it's really key clinically that we need to be able to risk adjust and show that it's working.

I'm just proposing those as potential

explanations. I obviously cannot speak for the people who changed, but that's just some potential explanations.

DR. RASTOGI: And just to feed back, exactly the same risk factors are used in all our models.

It's a software thing, you know. So, those things exist.

CO-CHAIR KOTTKE: So, whether or not the level changes much with the risk adjustment has nothing to do with the effectiveness of the risk adjustment.

I mean, there's no impact of the risk factors on risk even though it accounts for all of the risk. It's --- it doesn't --- it's not a criterion.

CO-CHAIR KOTTKE: Okay. So, should we move on? Thank you very much for your time this morning and this afternoon.

0067 ACC, chronic stable coronary artery disease, antiplatelet therapy.

MEMBER BRIGGS: So, this is a process

measure. And it's also a measure that's being brought to us --- is brought to us for maintenance.

It was initially endorsed in 2009. Re-endorsed in ---

CO-CHAIR KOTTKE: Excuse me just a second.

MEMBER BRIGGS: Oh, I'm sorry.

CO-CHAIR KOTTKE: We need the developers to be able to weigh in for a moment here. Welcome.

DR. HEIDENREICH: Sorry. Again, I'm Paul Heidenreich. I'm Chair of the Task Force for Performance Measures for the American College of Cardiology and the American Heart Association.

And for this measure, it was developed in 2003 along with the Physician Consortium for Performance Improvement of the American Medical Association. It's been in use since.

And the task force that uses this relies on Class 1 ACC/AHA recommendations and has a strong conflict of interest policy.

The data we show shows there still is a

gap in this measure. And given the strong relationship with mortality and hospitalization and morbidity, we feel antiplatelets are still a strong --- should still be a strong focus for improving care.

The testing shows very high reliability. It's currently in use by our Pinnacle, the ACC's Pinnacle Registry, as well as CMS' Physician Quality Reporting System, or PQRS.

Now, we know there are other measures. I know we're not doing a best in class at this time, but we feel this provides --- is still important given its --- that it's based at the individual clinician level.

It's registry-based, as well as being evidence-based, and reliable and valid. Thank you.

MEMBER BRIGGS: So, as I mentioned before, this is a measure that's being brought to us for maintenance endorsement.

And we looked at the evidence. And the evidence for this particular indicator is high.

There are multiple guidelines.

There's like four guidelines and like ten different statements on those guidelines that support this. So we felt like the evidence is high.

MEMBER VIDOVICH: Yes, no question. This is very high level of evidence. Hundreds of thousands of patients in multiple studies.

CO-CHAIR KOTTKE: Okay. So, anybody have any comments that they wish to make before we vote on the evidence?

(No comments.)

CO-CHAIR KOTTKE: Seeing no movement, let's vote on the evidence for antiplatelet therapy, 0067.

MS. IBRAGIMOVA: Importance to measure and report; 1a, evidence, structure, process, intermediate outcome. One, high, only eligible QQC submitted. Two, moderate. Three, low. Four, insufficient.

(Voting.)

MS. IBRAGIMOVA: Okay. The results are 12 votes for high. Zero votes for moderate.

Zero votes for low. Zero votes for insufficient.

CO-CHAIR KOTTKE: Opportunity for improvement.

MEMBER BRIGGS: So, we would agree that there is a performance gap in that the numbers are --- and the mean numbers are in the 80s, about 86 percent, but they've been pretty static. If you looked at 2009, it was 84.9 percent. In 2013, 86.2. And 2014, 86.3.

And while we feel it's an important measure, there is a question of whether we're topped out and what's the reason for no continued improvement in this particular indicator.

MEMBER VIDOVICH: In our preliminary discussions, one thing that I would like to point out is that patients --- it's the definition ---- what's the meaning of as-is chronic stable disease.

Are they by accident capturing some patients who underwent PCI and now are on aspirin and a bit of a -- antiplatelet therapy and might be counted as CAD, you know, and -- to help inflate the numbers. Like, you know so it looks better than

what it really is.

And on the other hand, could they be capturing some of the patients who are getting aspirin for another reason, let's say, a TIA or, let's say, 325 of aspirin for a different reason just maybe inaccurately given as the absolute number.

It probably is not a lot of patients, but I think there's probably some overflow here. Let's say somebody has a CAD, and then they forget to give them aspirin, and then they re-infarcted, and then they'll end up on aspirin and Plavix because now have a stent. And then they get counted as receiving aspirin.

So, just a minor issue. It's probably not a large proportion of patients.

CO-CHAIR KOTTKE: Gerry.

MEMBER MARTIN: First, I was going to say maybe we should have the developers answer the question, but then I was thinking that one of the things that does change over time particularly with some registries, is that you continue to enter in

new practices, new providers.

And so, what you don't have in this mean data is what's happening to individual, even, groups or practices.

So, it may be true that the countrywide is 86, but people that have been in it longer maybe they have improved and that data isn't shown.

CO-CHAIR KOTTKE: Yes. Mary?

CO-CHAIR GEORGE: I'd just like to add that 86 percent in this Pinnacle Registry seems really good. But when we look at what's happening across the country in our other major data sets at the population level, it's about 50 percent.

DR. HEIDENREICH: I'll also say regarding the PCI or even acute MI, those all, by definition, put you in the category of chronic coronary disease. You can never get out of that condition once you get into it.

And so, we'd say regardless of the reason you ended up on the right therapy, we would still give credit for being on the right therapy.

MEMBER BRIGGS: One of the issues I

think that we see is that, with aspirin, because it's a not prescription drug, that many times people don't consider that when someone asks them what their medications are in a medication list. So, it may not be documented by the staff in the office or the provider may overlook that as a possibility.

Obviously if you know that you're being monitored for that, you're more sensitive to that indicator, as in if you're dealing with a clinical registry, that kind of thing, but one possible answer for why we're seeing less than perfect scores in something that we think is so important is that aspirin is not considered on the same plane by some of the population, actually, in terms of when they think about drugs that they're on.

When somebody asks them about their medications many times, because it's not a prescription, they don't think about it that way.

CO-CHAIR GEORGE: We actually see a little bit higher rates in the patient report compared to the physician office reporting, but it is still sub-optimal.

MEMBER VIDOVICH: Just to bring -- Sometimes what happens is that the VAs, you know, the co-pay is the same whether somebody is on a brand or an aspirin. It's fixed rate.

And so, a lot of veterans just choose to buy their own aspirin and to take it with a co-pay from the VA pharmacy.

And then you can just slip through the cracks and it may not end up being recorded and state tracking is inaccurate and then maybe not be captured, but, again, minor issues.

I'm sure it's there in the finding.

CO-CHAIR KOTTKE: So, opportunity for improvement. Can we vote? Let's vote.

MS. IBRAGIMOVA: Importance to measure on report; 1b, performance gap. One, high. Two, moderate. Three, low. Four, insufficient.

(Voting.)

MS. IBRAGIMOVA: And the results are five votes for high. Seven votes for moderate. Zero votes for low. Zero votes for insufficient.

CO-CHAIR KOTTKE: Specifications and

reliability testing.

MEMBER BRIGGS: In terms of specifications, we have a question about the ICD-9 codes that include acute MI, because this is supposed to be an indicator for chronic stable angina.

So, again, if we're measuring --- if the standard is for angina patients and it's supposed to be stable angina, why are we coding in the specifications for acute MI?

DR. HEIDENREICH: Yes, the standard is actually not --- it's not angina. It would be chronic coronary disease.

And once you've had an acute MI, you then are, by definition, have coronary disease the rest of your life.

So, having acute --- looking for past MIs or even current MIs would be one way of identifying those with coronary disease.

CO-CHAIR KOTTKE: So, in fact, if you want to use the expression, it's overkill rather than missing. So, there's no crime. Any further

comments on reliability?

MEMBER BRIGGS: So for reliability they did signal to noise, and there was a very good result there at 0.994.

CO-CHAIR KOTTKE: Okay. Let's vote on reliability, please.

MS. IBRAGIMOVA: Scientific acceptability of measure properties; 2a, reliability. One, high. Two, moderate. Three, low. Four, insufficient.

(Voting.)

MS. IBRAGIMOVA: And the results are eight votes for high. Four votes for moderate. Zero votes for low. Zero votes for insufficient.

CO-CHAIR KOTTKE: Validity.

MEMBER BRIGGS: There was content validity done by expert work group, public comment, formal peer review process with the ACC Board of Trustees and Advisory Committees.

There's also construct validity done and face validity. Two different committees, one from ACC, and one from AHA, with 83 percent of them

agreeing that the measure is an accurate reflection of quality and being able to distinguish between poor and good quality.

Importance of the measure was rated 4.26 out of five.

CO-CHAIR KOTTKE: Anybody have any other comments?

(No comments.)

CO-CHAIR KOTTKE: Okay. Let's vote on validity.

MS. IBRAGIMOVA: Scientific acceptability of measure properties; 2b, validity. One, high. Two, moderate. Three, low. Four, insufficient.

(Voting.)

MS. IBRAGIMOVA: And the results are eight votes for high. Four votes for moderate. Zero votes for low. Zero votes for insufficient.

CO-CHAIR KOTTKE: Feasibility.

MEMBER BRIGGS: This measure is in use currently. So, I would say that it's feasible. It's a pinnacle data set, it's the one that's being

used --- registry.

CO-CHAIR KOTTKE: Any other comments?

(No comments.)

CO-CHAIR KOTTKE: Seeing no reaction, let's vote on feasibility.

MS. IBRAGIMOVA: Feasibility. One, high. Two, moderate. Three, low. Four, insufficient.

(Voting.)

MS. IBRAGIMOVA: And the results are ten votes for high. Two votes for moderate. Zero votes for low. Zero votes for insufficient.

CO-CHAIR KOTTKE: Usability and use.

MEMBER BRIGGS: Basically the same rationale. It's the pinnacle registry, is the current measure use. It's also being used and reported in the Physician Quality Reporting System.

CO-CHAIR KOTTKE: Other comments.

(No comments.)

CO-CHAIR KOTTKE: Seeing no reaction, let's vote on usability and use.

MS. IBRAGIMOVA: Usability and use.

One high. Two, moderate. Three, low. Four, insufficient information.

(Voting.)

MS. IBRAGIMOVA: And the results are 12 votes for high. Zero votes for moderate. Zero votes for low. Zero votes for insufficient information.

CO-CHAIR KOTTKE: Okay, so it's time to vote on whether to recommend the measure as suitable for endorsement.

This is 0067, chronic stable coronary artery disease, antiplatelet therapy. It seems to have cruised through. Linda got off easy.

Time to vote.

MS. IBRAGIMOVA: Overall suitability for endorsement. Does the measure meet NQF criteria for endorsement? One, yes. Two, no.

(Voting.)

MS. IBRAGIMOVA: We're just waiting for one more vote.

(Voting.)

MS. IBRAGIMOVA: And the results are 12

votes for yes. Zero votes for no.

(Pause.)

(Comments off the record.)

CO-CHAIR GEORGE: So, we have two measures left and we have just a quorum with what we have. So, we'll go ahead with 0079.

Developers?

DR. HEIDENREICH: Yes, so this measure on left ventricular ejection fraction assessment was developed also in 2003 on the ACC/AHA and the PCPI of the American Medical Association. And has also been used in the Pinnacle registry.

The importance comes from the fact that it is a requirement, in order to see if someone is a candidate for other performance measures already improved by NQF such as ACE inhibitors, beta blockers for heart failure that you have to have the ejection fraction to know if the patient can receive the --- will benefit from those.

There also is a gap currently in outpatient care. And you may be aware that Medicare dropped their --- I'll say their CMS core

measure because it was topped out in hospital, but that we are --- we believe our evidence shows for outpatient that is not the case, and the measure still has the utility in the outpatient setting.

In fact, you'll see only about 70 --- potentially only 70 percent is currently in use. We feel that testing shows high reliability and validity. And, again, it is currently in use in the ACC Pinnacle Registry.

CO-CHAIR GEORGE: Linda and Tom James.

MEMBER BRIGGS: Tom, are you there? Do you want to take this, or do you want me to take it?

MS. VICALE: Tom had to step out for a meeting between 2:00 and 3:00 p.m. Yes, so he said to refer to his comments that he provided in the worksheet.

(Pause.)

MEMBER BRIGGS: So, again, it's a process measure. It's looking at people 18 years of age or older with heart failure who have a quantitative or qualitative result of a recent prior LVEF documented within 12 months. However,

recent is not really defined within the measure.

We all know that it's really important to have that to base therapies from. There's no argument there.

But as I said, the numerator is within --- documented within the last 12 months, but it doesn't necessarily mean that it was done within that 12-month time period.

And the documentation that --- it uses a registry, and the registry is the Pinnacle Registry again. And you can either have a number, or it can be in a range. So, on the Pinnacle Registry form you can either --- there's a blank and you can fill in the number for the ejection fraction, or you can pick something that's hyperdynamic, greater than 70, normal being 50 to 70, mild dysfunction, 40 to 49. And so, there are some instructions to how to put it in that frame.

And then I guess you could also --- because it refers to documentation. It could be in the note that way, potentially.

So, in terms of the evidence for this,

the evidence is only by expert opinion from the heart failure guidelines.

CO-CHAIR GEORGE: Any comments on the evidence?

(No comments.)

CO-CHAIR GEORGE: So, how would you rate that if it's expert opinion only?

MEMBER BRIGGS: I would have to say that it's with exception. So, it's insufficient with exception.

DR. HEIDENREICH: I'll say while you're thinking, that while it is a Class C, I don't think there has been or ever was even considered the possibility of a randomized trial given that all patients with heart failure -- if you're going to provide life-prolonging therapy, you have to get the ejection fraction.

So, there never was the thought that you would not get an ejection fraction in patients with heart ---

MEMBER BRIGGS: I would agree with that, but I just have to go by what the --- what our

rating system is.

MEMBER AL-KHATIB: What I would add to that is that there is a very clear association between measuring the ejection fraction and the outcomes. And then, as was stated by the developer, in forming the treatment plan that you come up for the patient. So, I think the evidence is pretty strong.

CO-CHAIR KOTTKE: But inadequate with the exception, I mean, it's not a chronic.

MEMBER BRIGGS: No.

CO-CHAIR KOTTKE: It's just a statement of fact that this is a different type of issue that, I mean, I agree with you that --- and I think that it's unconscionable that EMRs can't come up with ejection fractions, but I think it goes ahead if we say it's -- you know, there's an exception.

CO-CHAIR GEORGE: I think that's the way the algorithm works. And we just need to have that option for voting, which we don't have on the screen.

MR. CHIU: The only thing -- if I can

add, Dr. Kottke, really quickly, is I think there is a comment about most recent, you know. I think we try to get it from the past year or two, but it really is --- I think somebody asked in the comments to any time in the past.

And the point is if you're really low AEF, 20 percent, 30 percent, other measures we have in our other partnerships of antiplatelet and other measures, ACE and beta blocker, they all relate to those kind of measures.

Hence, we thought that if it's a low rate at AEF, you don't need to continually check it.

MEMBER BRIGGS: I totally agree with the fact that this is important, but I just have to go by what the algorithm is.

(Pause.)

MS. WILBON: Sorry. We're just conversing on the use of the evidence exception. We just want to make sure we're using it appropriately.

So, with the algorithm, if you look at the orange boxes where it says, are there or could

there be performance measures of related health outcome or evidence-based intermediate clinical outcomes of process.

So, if the committee agrees that there could be another measure that would be closer to the outcome, then you would not apply the exception.

But if you believe that there would not be and there was expert opinion, then you could go on to ---

MEMBER BRIGGS: That's basically where we're at, yes.

MS. WILBON: Yes, I just wanted to make sure that we're accepting expert opinion because we don't think there will ever be a study to look at the utility of ejection fraction in these patients.

It's kind of -- everybody needs that information, uses that information in all the studies, base treatment on those things. So, it's kind of the bedrock --- kind of a bedrock kind of thing.

(Comments off the record.)

MS. WILBON: Yes, let's go ahead and

vote on this. And then we'll discuss the next steps once everyone has submitted their vote, because you can't --- you can't automatically apply the exception until we get through this vote.

DR. JOHNSON: And let me be clear, what you would do here is if you feel like you're going to land on asking for the exception, vote here for insufficient. We'll have another vote exception yes or no. Okay. So, we're splitting it into two votes.

MS. WILBON: Yes.

DR. JOHNSON: Thank you.

CO-CHAIR GEORGE: So, I'll just summarize. We are going to vote on the evidence. It's expert opinion.

Because of the way the slide is, we can't add the insufficient with exception option on this slide.

If that is where we're headed, we would vote insufficient on this slide. And then we will vote yes or no to add with exception on a second vote.

(Comments off the record.)

CO-CHAIR GEORGE: Any questions on --- okay. So, we'll go ahead and vote on the evidence.

MS. IBRAGIMOVA: Importance to measure and report; 1a, evidence, structure, process, intermediate outcome. One, high, only eligible QQC submitted. Two, moderate. Three, low. Four, insufficient.

(Voting.)

MS. IBRAGIMOVA: And the results are one vote for high. Zero votes for moderate. Zero votes for low. Eleven votes for insufficient.

CO-CHAIR GEORGE: So we'll now vote on whether insufficient alone or insufficient with exception, yes or no --- or, it's insufficient with exception, yes or no. What is it?

MS. WILBON: So, we just want to just make a clarification. So, on Box 10 --- and if I missed this discussion, forgive me, but I just want to make sure that we're clear on the use of the exception that the committee has considered. What we kind of ask you guys to consider in the evidence

criterion.

So, are there --- or there could be performance measures of a related health outcome or evidence-based intermediate clinical outcome or process.

So, this is a process measure looking at the assessment of the ejection fraction for patients with heart failure.

So, the question would be, is there a different --- could there be a different measure that is closer to what you would see would be the outcome that is an intermediate outcome or outcome that would better --- a measure that would get us closer to the outcome. Could there be something there?

And if not, then, you know, we can move on. But I just want to make sure that we have that conversation, because I feel like we kind of jumped to the evidence exception.

The other thing about the evidence is it doesn't have to be the RCTs. There's other types of studies and other types of evidence that could

also be considered. There's a gap between expert opinion and RCTs. So I just wanted to make sure that we're considering that body as well.

MEMBER BRIGGS: So I would say that this, again, is a very key measure. It's something that we use to base other therapies on.

So, I don't see that there's anything that's going to replace it. And I don't see that an outcome necessarily is going to be better than this particular indicator that we do need the ejection fraction to make clinical decisions on our heart failure patients.

So, I think that it's important from that regard. I don't think that there's anything else that you're going to find that can replace that, because basically in all the tests that we do, we're doing a cath, we do a ventricular-gram to get the EF. We do an MR, we're trying to get the EF. We do an echo, we're trying to get the ejection fraction. We're looking for that data.

Okay? So, and then this is all from -- the expert opinion is from clinical guidelines and

it's been echoed over and over again. So, I think that that's a good basis to work from.

CO-CHAIR KOTTKE: Right. And so, in Box 12 if we feel that there is --- benefit outweighs harm, we vote for exception.

MEMBER AL-KHATIB: So, just a quick comment. I appreciate you are reminding us that we don't need randomized clinical trial data, but there are a lot of epidemiologic data that correlate the EF with outcomes.

Does that not qualify? I'm confused now, actually.

MS. WILBON: It does. I know she made a statement earlier about there's never going to be ---

MEMBER AL-KHATIB: A randomized clinical trial.

MS. WILBON: Right, so I was just responding to that.

MEMBER AL-KHATIB: But we have a lot of observational data that have proven the usefulness and the importance of this.

So, does this elevate it to the important -- like that there's evidence, the high evidence, or are we still talking about the exception, I guess is my question.

CO-CHAIR KOTTKE: But there is no --- is there evidence that just measuring improves outcomes? I mean, it changes behavior ---

MEMBER AL-KHATIB: Indirectly, yes.

CO-CHAIR KOTTKE: -- but measurement alone doesn't change outcomes.

MEMBER AL-KHATIB: Of course it does. Because if you don't get an EF, you won't be able to know that the patient needs an NICD, for example. And that ---

CO-CHAIR KOTTKE: No, just measuring and ---

MEMBER AL-KHATIB: -- saves lives.

CO-CHAIR KOTTKE: -- doing nothing else does not improve outcomes.

MEMBER AL-KHATIB: But I don't know that you can separate the two. I'm not sure that I follow.

CO-CHAIR KOTTKE: I'm sure you can.

MEMBER AL-KHATIB: No, I don't --- I don't see that. As clinicians, we don't just get a test and not act on, you know, upon the result of the test.

That's how we function as clinicians. We get a test and we do something with it.

CO-CHAIR KOTTKE: No, you need the test to make the decision, but getting the test alone doesn't help the patient if you don't do anything with it.

MEMBER AL-KHATIB: I don't know. I'm confused.

CO-CHAIR KOTTKE: I'm not saying it's appropriate. I'm just saying the test alone doesn't help.

CO-CHAIR GEORGE: Ashlie, I have a question.

MS. WILBON: Yes. Sure.

CO-CHAIR GEORGE: Are you asking us to vote on the question in Box 10, or the question in Box 12?

MS. WILBON: So, they go hand in hand.

CO-CHAIR GEORGE: But it makes a difference whether you answer yes -- whether you answer no to 10, or yes to 12.

MS. WILBON: Right. So you don't actually get --- you don't actually get to vote on the exception unless you voted that the evidence that they submitted was insufficient, which it sounded like based on the committee's discussion that they felt that the evidence that was submitted in the form from the developer was insufficient. So, that's ---

CO-CHAIR GEORGE: What I'm trying to clarify is what is a yes vote and what is a no vote.

MS. WILBON: On the exception?

CO-CHAIR GEORGE: On what we're going to vote on now, yes.

MS. WILBON: Oh, okay. So you believe that although the evidence was insufficient, that there should be an exception applied that would allow the measure to go forward.

CO-CHAIR GEORGE: That would be a yes

vote if that's what we believe.

MS. WILBON: Yes.

CO-CHAIR GEORGE: Okay. I just want to clarify that.

MS. WILBON: Yes.

MEMBER AL-KHATIB: Can we re-vote on the evidence? Based on this discussion, I'd like to change my vote. Is that possible?

MS. WILBON: Yes.

MEMBER AL-KHATIB: Because, I mean, again clinically there is not a single test that we do that changes outcomes. It's whatever we do with the result of the test that changes outcomes.

I get a patient with syncopy. I put them on a monitor. They have a nine-second pause. I put a pacemaker in them. That's what's going to make them live longer. It's what I do with the data.

You can't separate the two clinically. It's impossible to separate. Then not a single test performance measure that's based on a test will pass the evidence criterion. Not a single one.

Because just doing the test per se is not going to do anything. It's not going to change anything. It's how you use the data to manage the patient.

DR. JOHNSON: So, let me try to clarify a little bit. That's exactly why we have that question in the evidence exception, the Box 10, because the idea is that there can be lots of things that are important to do in practice, right? And testing and assessing and those kind of things are important. They're the first step to getting somewhere, but then you have to act on it and do something about that.

So, there is lots of different kinds of measures that you could build. So, you could build measures just about the assessment, because that's important, or you could build measures further down the line and talk about a treatment or something like that or the actual outcome.

So, NQF actually has a hierarchy of preference that we prefer to endorse the measures that are closer to the outcome.

So, that's why we have this question

here. So it's not saying that it's not important to do those things. The question is, is it important to have a national consistent standard that basically that's just out in the world as an NQF-endorsed measure. So, that's the question.

MEMBER SPANGLER: Can I add something real quick? Sana, I'm not disagreeing with you about the epidemiological evidence, but I don't think that was submitted with this.

And according to the algorithm it says there's empirical evidence submitted. So, it might be there, but it wasn't in the application. So, I think that seems like why we need to go down to that orange path.

MEMBER BRIGGS: There were two studies submitted with the evidence, but one study was actually on an entirely different topic related to CMR.

Obviously you can get an ejection fraction that way, but it was more about the CMR than it was about the ejection fraction itself.

And I'm not finding the other one, but

neither one of them were directly related to the ejection fraction and the use of it on patients.

MEMBER DELONG: Well, I think we're splitting hairs here. And if we only have a half hour to cover the next one, it's obvious how the vote is going to turn out.

CO-CHAIR GEORGE: All right. Any last questions before we vote?

(No comments.)

CO-CHAIR GEORGE: All right.

MS. IBRAGIMOVA: So, can we first agree that the question on the screen is the voting question?

Should there be an exception applied that would allow the measure to move forward? One yes. Two, no.

(Voting.)

MS. IBRAGIMOVA: So, the results are 12 votes for yes. Zero votes for no.

CO-CHAIR GEORGE: Let's quickly move on to opportunities for improvement.

MEMBER BRIGGS: So, there was a

significant performance gap for this measure.

In 2013, the mean compliance with this was 67 percent. In 2014, 72.5. So, there's obviously room for improvement in terms of documentation of this in the outpatient environment, which is what this particular one is about.

And obviously if we're using this for a critical basis for our treatment, it's important.

CO-CHAIR GEORGE: Any comments on the opportunity? If not, we'll vote.

MS. IBRAGIMOVA: Importance to measure on report; 1b, performance gap. One, high. Two, moderate. Three, low. Four, insufficient.

(Voting.)

MEMBER CHO: Can I just say how I think that number is so low in the ear of overtesting in America, how is it possible?

Because I have patients who are referred -- I work at the Cleveland Clinic -- who are referred to me and they get an echo or a stress test every six months.

That number of 67 percent I find almost improbable.

CO-CHAIR KOTTKE: (Speaking off mic.)

MEMBER CHO: No, I actually think you're wrong, because we're not the ones overtesting. These are people being referred to us from whatever many states or places around the country.

And there are plenty of people getting -- you -- I don't need to tell you, are getting overtesting.

CO-CHAIR KOTTKE: (Speaking off mic.)

MEMBER BRIGGS: I didn't mention that there are disparities, but it's by actually insurance that we have information.

And actually, believe it or not, the Medicare population did the worse in terms of having this reported.

If they had no insurance at all, it was reported 69 percent of the time. Private insurance, 64.4. Medicaid, 60 percent. And Medicare, 47.6 percent.

DR. HEIDENREICH: One error in that is that the Medicare and Medicaid lines were switched.

MEMBER BRIGGS: Oh, really? Okay. Well, that makes sense.

DR. HEIDENREICH: But there clearly are differences by insurance. That doesn't take away that there are differences in insurance.

MEMBER BRIGGS: Yeah. No, there are disparities for sure.

MS. IBRAGIMOVA: So, the results are 11 votes for high. One vote for moderate. Zero votes for low. Zero votes for insufficient.

CO-CHAIR GEORGE: Move on to reliability and specifications.

MEMBER BRIGGS: So, this was a reliability test in using the Pinnacle data. They used 2,254 providers and 409,000 plus patients. And it was a signal to noise and very good numbers.

If they -- if providers had more than ten patients, the average reliability was .988 in 2013. And .989 in 2014.

CO-CHAIR GEORGE: Any comments on the

reliability or specifications?

(No comments.)

CO-CHAIR GEORGE: If no, we'll vote.

MS. IBRAGIMOVA: Scientific acceptability of measure properties; 2b, reliability. One, high. Two, moderate. Three, low. Four, insufficient.

(Voting.)

MS. IBRAGIMOVA: The results are 12 votes for high.

CO-CHAIR GEORGE: Validity.

MEMBER BRIGGS: They did content validity and face validity. The content validity was assessed with the expert work group. It was not for public comment.

Also in formal peer review processes and the ACC Board of Trustees also assessed this. And the PCPI membership as well.

So, there was construct validity also and face validity was looked at by two committees. One by -- from the ACC, and one from AHA with 87 percent of -- 85 percent of the members agreeing

that the measure was an adequate reflection of quality and was able to distinguish between poor and good.

And in terms of importance it was rated 4.24 out of five.

CO-CHAIR GEORGE: Any comments on the validity?

(No comments.)

CO-CHAIR GEORGE: We'll vote.

MS. IBRAGIMOVA: Scientific acceptability on measure properties; 2b, validity. One, high. Two, moderate. Three, low. Four, insufficient.

(Voting.)

MS. IBRAGIMOVA: And the results are 10 votes for high. Two votes for moderate.

CO-CHAIR GEORGE: Feasibility.

MEMBER BRIGGS: This is a current measure. It's being used in the Pinnacle Registry. Also, there's abstraction through MDS and OASIS for that. So, I would say it's feasible.

CO-CHAIR GEORGE: Any comments on

feasibility?

(No comments.)

CO-CHAIR GEORGE: We'll vote.

MS. IBRAGIMOVA: Feasibility. One, high. Two, moderate. Three, low. Four, insufficient.

(Voting.)

MS. IBRAGIMOVA: The results are ten votes for high. Two votes for moderate.

CO-CHAIR GEORGE: Usability.

MEMBER BRIGGS: Again, the developers site the Pinnacle Registry and they continue to seek opportunities for public reporting. However, currently it's not publicly reported.

CO-CHAIR GEORGE: Any comments on usability?

(No comments.)

CO-CHAIR GEORGE: We'll vote.

MS. MARINELARENA: Before we vote, I just want to note that the NQF policy states that measures -- this has been endorsed since 2009. And after six years, they should be publicly reported.

And this right now is not publicly reported.

DR. HEIDENREICH: I'll say it was publicly reported, I think, through PQRS -- or I don't know if PQRS is considered public reporting.

It was publicly collected, but they have -- in terms of outpatient, they have not publicly reported outpatient ones.

MS. SPEAKER: It got removed from PQRS before they started doing the public reporting for the ACL Limited Measure stat. But, you know, we have found that this is a very important measure from a quality improvement perspective for the reasons that have been mentioned before, which is if you don't do it, you can't figure out what type of therapy they need whether it's medication, device or whatever.

CO-CHAIR KOTTKE: And is that a recommendation or a strict measure of the NQF?

MS. SLATTERY: You know, as we have stated, I think, in previous phases for the project, the ACC has not started publicly reporting either at the practice or physician level, in part, because

we do use these measures for submission to PQRS.

We do see shifts that will tap in with this measure, and then vis-a-vis Pinnacle's ability to qualify as what's called a qualified clinical data registry for PQRS submission requirements.

We have been waiting to see what is going to happen in the regulations so that we are not duplicating our resource investments to publicly report at either the practice or physician level if it's going to have to be publicly reported on Physician Compare.

Some recent changes to PQRS and giving QCDRs the option to either let the physicians use Physician Compare as a publicly reporting mechanism, or for a specialties society to create a public reporting mechanism are now being considered by ACC, but it is a very expensive enterprise for us to be able to engage in public reporting.

And I think that there are regulatory changes that have been evolving that have -- that the college has had to consider in weighing out

whether to invest in developing a public reporting program.

CO-CHAIR KOTTKE: So, the answer from NQF is what?

MS. WILBON: So, I'll give a short statement, which is that the purpose of NQF endorsement is to endorse measures for quality improvement and accountability applications.

CO-CHAIR KOTTKE: Right.

MS. WILBON: So, the whole point of us giving developers six years to have their measures publicly reported is so that they could meet that accountability -- I won't say criterion, but that level of which endorsement is really -- the level of endorsement is really intended to be measures not just that are being used for QI, but that are being used in accountability applications. One of which is publicly reporting.

So, I would just say that, you know, the intent of an NQF endorsement is for measures that are being used for those purposes.

CO-CHAIR KOTTKE: Well, let me ask you

a different question.

MS. WILBON: And if not's being used for that, then --

CO-CHAIR KOTTKE: Can we endorse it if it is not publicly reported after six years?

MS. WILBON: I believe the committee can certainly recommend endorsement.

CO-CHAIR KOTTKE: Okay. Good.

MS. WILBON: There's a lot of process left, but --

CO-CHAIR KOTTKE: Let's move along.

MS. SPEAKER: So, it wasn't publicly reported because when it was in PQRS, it was before they started doing that one. We're figuring out did they want it physician level, or physician group practice level.

So, they barely started public reporting about a year and change ago. And I think that, you know, as Laura mentioned, we're going to figure out their website, our website --

CO-CHAIR KOTTKE: So, my read is that we can endorse it without even though they haven't

publicly reported it. And so, I think we ought to move along.

MEMBER DELONG: Kristi had some -- I was actually interested in hearing what you had to say.

MEMBER MITCHELL: I was going to actually respond to what Laura was talking about regarding QCDR. So, I'm good, but I am -- this whole thing does bring up a bigger point regarding the ultimate intent of these measures to be used for accountability rather than quality improvement.

And I guess, you know, I must have missed this piece in the memo about six years -- having six years to move a measure from QI to accountability.

Not all measures should be intended for accountability.

MR. WILBON: No, I'm sorry. That was misinterpreted. For QI and accountability purposes, but measures -- we generally don't endorse measures that are just for QI.

So, just to -- it could be used for QI. Not to say that it couldn't. Once it's an accountability application that it can't be used

for QI, but Karen might be able to clarify things.

DR. JOHNSON: Right. So, we do right now endorse measures for both purposes. So, the idea is the measures should be suitable for both even if you're not in both.

I will point out that our guidance for the usability criterion actually asks for use and accountability program within three years. And if it's in like PQRS or something like that, then it hits that mark.

The six years is public reporting. That's another kind of accountability program. And it's really getting to the desire to let consumers and purchasers and the public understand measure results. So, it's kind of going one step further.

Usability and use, number one, we don't have absolute thresholds on anything. You probably have noticed that with all of our discussions about reliability and validity and that sort of thing, but we do have the guidance.

That said, usability and use is also not

unless past criteria.

CO-CHAIR KOTTKE: So, can we -- I am leaving in 17 minutes. The lights go out in 17 minutes. We have one more --

MS. VICALE: Can I just add that Tom James is set to come back at 3:00 p.m.?

CO-CHAIR KOTTKE: Okay.

MS. VICALE: So, that would keep us at quorum.

CO-CHAIR KOTTKE: Good. Good. I'm still leaving in 17 minutes.

(Laughter.)

CO-CHAIR KOTTKE: Okay. Can we vote on this?

MS. IBRAGIMOVA: Usability and use. One, high. Two, moderate. Three, low. Four, insufficient information.

(Voting.)

MS. IBRAGIMOVA: So, the results are six votes for high. Four votes for moderate. One vote for low. One vote for insufficient information.

CO-CHAIR GEORGE: All right. Any last-minute comments before we vote on the overall measure?

(No comments.)

CO-CHAIR GEORGE: Seeing none, we'll vote on the measure.

MS. IBRAGIMOVA: Overall suitability for endorsement. Does the measure meet NQF criteria for endorsement? One, yes. Two, no.

(Voting.)

MS. IBRAGIMOVA: The results are 12 votes for yes. Zero vote for no.

CO-CHAIR GEORGE: Thank you, developers.

MEMBER DELONG: While we're transitioning to the next measure, I think what we just discussed and the implication which confuses me as to what our endorsement means is not totally clear.

If our endorsement means this will be used for accountability later, I think we need to take that seriously and only endorse measures that

we feel will validly hold providers to an accountable level.

CO-CHAIR GEORGE: Do you think that should be a voting consideration criteria?

MEMBER DELONG: I don't know. I just see us endorsing measure after measure that -- for which the data aren't necessarily there, but we're saying it's the best we can do.

I would not want to be held accountable for a measure that didn't have the appropriate level of data quality and completeness.

CO-CHAIR GEORGE: Just a few brief comments, very brief, from the developers.

DR. KAUFMAN: Good afternoon. We thank the NQF for this opportunity to appear today as WCHQ seeks NQF endorsement for our all-or-none IVD, ischemic vascular disease measure.

I am here with Mary Gordon who is clinical information manager at WCHQ. I'm a general internist, former chief medical officer at Dean Clinic in Madison, Wisconsin and serve as WCHQ's clinical advisor.

WCHQ was founded in 2003. It's a voluntary membership-driven organization dedicated to public performance reporting, shared learning in order to improve the quality of care and affordability of healthcare in Wisconsin.

The membership includes 38 healthcare organizations and those organizations care for more than 65 percent of Wisconsin citizens.

Members actively use our 44 publicly reported measures to drive internal improvement efforts.

Our all-or-none IVD outcome measure has four individual components; blood pressure control, aspirin or other antiplatelet medication use, tobacco-free status and use of a statin medication.

The measure is consistent with the 2011 AHA/ACC foundation guideline for secondary prevention of atherosclerotic heart disease -- vascular disease and the 2013 ACC/AHA task force guideline on treatment of blood cholesterol.

WCHQ began reporting some of the

individual IVD component metrics in 2012 and first reported the all-or-none IVD measure in the spring of 2015 for over 50,000 patients.

Current performance in the all-or-none measure ranges from 45 percent to 70 percent. We believe that the all-or-none measure methodology compared to only reporting on the four component metrics individually provides a more comprehensive view of the care provided, is a more sensitive indicator of care quality, will be a greater spur to organizational improvement and will make it easier for the public to understand differential performance among our membership.

We appreciate the opportunity to appear today and look forward to answering questions. Thank you.

CO-CHAIR GEORGE: Sana or Leslie.

MEMBER AL-KHATIB: Yes, of course. Thank you very much for that overview. We're going to do our best to finish this in ten minutes, or close to that anyway.

So, I think this measure is very clearly

described. I just have a couple of questions for you and then we can delve into the voting unless other people have questions.

So, in terms of the documentation in the medical record of statin use, do you allow for contraindications and is that captured?

MS. GORDON: Currently we do not do that, but we have discussed that. We have a -- what we call our Ambulatory Care Specifications Committee. And we also have a Measurement Advisory Committee.

We did talk about allowing contraindications. And at this point in time with -- we did not opt to do that because there wasn't, you know, like a clear ICD-9 or even ICD-10 code. And we just weren't sure how we were going to capture that adequately electronically.

I guess that's the other piece of it is that at least as far as the WCHQ membership goes, we kind of strive to have our measures be electronically, you know, the data to the industry feels that we can capture electronically.

DR. KAUFMAN: It's a really challenging issue because there aren't specific ICD-9 codes, myalgias or whatever --

MS. GORDON: Right.

DR. KAUFMAN: -- that you can directly correlate that it's a statin medication use. We would love to get there, but it's hard. ICD-9 really isn't specific enough right now.

MEMBER AL-KHATIB: I hear that. Thank you. And then one other question from me and we'll see if you guys have any questions.

When you talk about CAD risk equivalent condition, I mean, are you referring to PAD and can you elaborate a bit more on that?

DR. KAUFMAN: I think it's consistent with the 2011 guideline. Coronary artery disease, atherosclerotic vascular disease, including peripheral artery disease --

MEMBER AL-KHATIB: Okay.

DR. KAUFMAN: -- aortic disease and carotid artery disease. So, I think we have structured it so it's exactly the same.

MEMBER AL-KHATIB: Okay.

MEMBER CHO: Diabetes.

DR. KAUFMAN: Diabetes, per se, without vascular disease?

MEMBER CHO: Diabetes as considered a coronary artery disease equivalent.

DR. KAUFMAN: Diabetes alone is not included right now.

MS. GORDON: Not diabetes alone currently.

MEMBER CHO: Okay.

MEMBER AL-KHATIB: Any other questions before I start? I can start in -- so, talking about the evidence, I mean, clearly this measure addresses significant health problems.

The evidence is very clear regarding the association or really the causality between all of these component factors of this measure and outcomes. So, I have no concerns about the evidence.

CO-CHAIR GEORGE: Was there a QQC or --

MS. SPEAKER: This is a composite

measure.

CO-CHAIR GEORGE: Any comments on the evidence?

(No comments.)

CO-CHAIR GEORGE: If not, we'll go ahead and vote.

MS. IBRAGIMOVA: Importance to measure on report; 1a, evidence, structure, process, intermediate outcome. One, high. Two, moderate. Three, low. Four, insufficient.

(Voting.)

MS. IBRAGIMOVA: And the results are 12 votes for high.

CO-CHAIR GEORGE: Thank you.

MEMBER AL-KHATIB: In terms of opportunity for improvement, the developers provided some compelling data showing that they actually tested 121 clinics covering a total of 42,290 patients. And they showed that the average clinic performance on the measure was .5862. And the range was .379 to .75.

So, with this variability and this gap,

I think it's a no-brainer.

CO-CHAIR GEORGE: Any comments on the performance gap?

(No comments.)

CO-CHAIR GEORGE: All right. We'll vote.

MS. IBRAGIMOVA: Importance to measure and report; 1b, performance gap. One, high. Two, moderate. Three, low. Four, insufficient.

(Voting.)

MS. IBRAGIMOVA: And the results are 12 votes for high.

MEMBER AL-KHATIB: So, for reliability I think we're talking, or is this the composite? The composite makes perfect sense to me. I have no concerns there.

CO-CHAIR GEORGE: Any comments?

MEMBER MITCHELL: I have a comment about the tobacco-free element of the composite. It's sort of you're reflecting on your experience with the D5 and the challenge relative to that single component getting it to a point that's

reliably collected, and validly assessed.

MS. GORDON: What we have been -- we have been -- WCHQ has been reporting tobacco status and tobacco association just on measure of sort of the global populations of our membership, I don't know, for quite a few years.

And I know one thing that we've seen in reporting that as far as the data being reliable is that what we had noticed when we first started reporting that measure was that we had some organizations that had very low results and, you know, it tended to be that they didn't have a process either in place that it wasn't happening, or it just wasn't being adequately documented.

And so, we saw great improvements just from publicly reporting this measure in our membership's results in that area. And we a couple of years ago moved on to a tobacco-free measure of diabetes, which would be the same measure that's incorporated here.

And we feel just from our experience with how it's progressed with our tobacco

association measure that I think we feel good about the data, you know, and the data collection methods that are being used.

MEMBER CHO: So, does the patient have to quit in order for it to -- you have to -- so, if you counsel a patient, you documented that you counseled a patient, but there is no quit date. And that would be a --

MS. GORDON: A fail for this particular measure, right. For the tobacco cessation measure, which is not part of this composite, there if they're just counseled, it counts as a numerator compliance, but for this one we're actually looking to see if the patient, you know, does not smoke.

DR. KAUFMAN: Right. They are tobacco-free if they are compliant in this measure. It's just not counseled. It's just not asked.

MEMBER AL-KHATIB: We realize we don't have full control over that, but, I mean, I certainly am not opposed to having it in the measure. We certainly don't have full control over that.

DR. KAUFMAN: I don't think we have total control over a lot of things as physicians.

CO-CHAIR GEORGE: Any other comments?

MS. WILBON: So, I -- of course me again. So, I realize that we're short on time. We're going to lose Tom shortly and we're not sure if the other Tom is going to be on the phone.

We just want to make sure --

MEMBER JAMES: I'm back.

MS. WILBON: Oh.

MEMBER JAMES: I just sent a note.

MS. WILBON: Okay. Yeah, so we just want to make sure that everyone is comfortable. There are some components -- I know the staff did a review of this and there are some components of the measure.

We just want to make sure that the committee does do, you know, due diligence discussing terms of the blood pressure parameters. I know we had some hypertension measures before where the parameters were discussed. In terms of dosages of medication, like, we just want to make

sure that there's adequate discussion in terms of consistency, because we find these things sometimes come back.

So, while we're all here gathered and we're discussing the measure, I just want to make sure that everyone, you know, that we do our due diligence and make sure that we're, you know, addressing all the different components of the measure. So, just --

MEMBER AL-KHATIB: Okay. From the blood pressure standpoint this is definitely in line with the guideline document that uses this cutoff. Certainly stopping smoking, I don't think anybody has to, you know, say much in relation to that.

Aspirin antiplatelet therapy, that's absolutely, you know, in there. Use of statin, I mean, I think all of these are very well supported by data as well as guideline recommendations.

CO-CHAIR GEORGE: Any other comments on the construct validity?

(No comments.)

CO-CHAIR GEORGE: If not, we'll vote.

MS. IBRAGIMOVA: Importance to measure and report; 1c, composite. One, high. Two, moderate. Three, low. Four, insufficient.

(Voting.)

MS. IBRAGIMOVA: So, the results are nine votes for high. Three votes for moderate. Zero votes for low. Zero votes for insufficient.

CO-CHAIR GEORGE: Specifications and reliability testing.

MEMBER AL-KHATIB: So, we covered a lot of the specifications because the numerator has to do with and meeting these, you know, goals, if you will, for blood pressure, tobacco use, antiplatelet use, statin use. So, no issues there. Denominator makes perfect sense to me. They had no exclusions.

In terms of testing, I think based on the data that they provided that the reliability is as good. They derived data from 17 group practice members of the Wisconsin Collaborative for Healthcare Quality.

15 groups reported all electronically. Two groups reported using random sample methodology. And they really got data from 121 sites covering more than 50,00 patients and provided reasonable data. Average reliability was found to be .7817.

And from my perspective, I find this acceptable.

CO-CHAIR GEORGE: Any comments on reliability?

(No comments.)

CO-CHAIR GEORGE: If not, we'll vote.

MEMBER JAMES: This is -- just one question. This is listed on the worksheet as not an e-Measure, but the data is really obtained primarily through electronic methods.

I'm presuming that we're starting this thing off not as an e-Measure, but it could be developed into one; could it not?

MS. GORDON: That's correct.

MEMBER JAMES: It would enhance reliability, I think.

MS. GORDON: Yes, that is correct. It's not currently developed as one, but it certainly could be.

MEMBER JAMES: Okay. Thank you.

CO-CHAIR GEORGE: All right. We'll vote on reliability.

MS. IBRAGIMOVA: Scientific acceptability of measure properties; 2a, reliability. One, high. Two, moderate. Three, low. Four, insufficient.

(Voting.)

MS. IBRAGIMOVA: The results are six votes for high. Six votes for moderate. Zero for low. Zero for insufficient.

CO-CHAIR GEORGE: Validity.

MEMBER AL-KHATIB: So, with regard to validity, what they did is they have assessed this measure for data validity. And they talk about, you know, how the measured numerator for each reporting entity is subject to validation once every three years on a schedule based on random selection.

And they talk about how the results that vary greatly between reporting periods or that appear significantly or higher or lower than the mean are subject to validation and I feel like the whole plan that was outlined for validity testing was pretty good.

I'm not sure that I was able to find where the numbers are, where the results of the testing were. I think that was a question I had.

CO-CHAIR GEORGE: Leslie.

MEMBER CHO: My other question is, is that in one part of the measure it talks about statin therapy. In another part of the measure it talks about high-intensity statin therapy unless it's contraindicated. I don't know which one we're testing the population in. That's my first question.

My second question is, is I still remain concerned about how you would exclude patients who are statin-intolerant, because it affects, you know, five to ten percent of the population, as we know.

DR. KAUFMAN: To your first question, it's simply are you on statin therapy or not?

(Comments off the record.)

DR. KAUFMAN: Okay. On Page 7.

MEMBER CHO: Okay.

DR. KAUFMAN: Okay. And then in terms of the intolerance, really, because so much is collected electronically and it's through ICD-9 or even 10 codes. You can't connect side effect ICD categories to the actual medication.

It's challenging at this point without doing record review on all the patients. And it's -- although it's a level playing field, obviously, for the collaborative and for improvement efforts, everybody's reported measures have the same issue. So, again, it's relative performance.

MEMBER CHO: But if you have a measure that we're not measuring correctly, we'll never be able to get the right -- like let's say you go from 70 to 80 and --

DR. KAUFMAN: Right.

MEMBER CHO: -- we never go above 80.

Well, maybe we never go above 80 because we're not capturing the statin-intolerant population.

DR. KAUFMAN: Right. And I think that's really a valid comment, but it's analogous to when the physicians at Dean used to come to me and talk about their diabetes measures, you know, and maybe a 95-year-old patient who is a diabetic who has terminal cancer doesn't need a hemoglobin A1c.

So, you know, it's not perfect. For sure it's not perfect. It would be great if we could correlate and capture that easily in a large database and that would be our goal down the road for sure, but it's challenging sometimes.

MEMBER AL-KHATIB: Go back to the question that I raised. Because as I said, you know, what I see in front of me is the outline of what you've done to validate, which I think makes perfect sense, but I don't see the data. I don't see the results.

Can you direct me to where those are?

MS. GORDON: I don't know that we

actually included any kind of numerical results only that we indicated that through the validation of this composite measure.

And it was partially also because we composed the statin use measure that we had two entities that we identified that did not publicly report the measure this first time around because of some issues that we found with their -- it was really related to the statins.

So, I -- other than that I don't really have -- there really wasn't numerical information --

MEMBER AL-KHATIB: Right.

MS. GORDON: -- included in this.

MEMBER SPANGLER: Mary, there is something on Page 59. Can you explain what that data is? It says, published results of the group level ranged from -- and gives numbers.

MS. GORDON: Page 59. Let's see. I'm not sure if I -- on Page 59, I'm not sure which document.

MEMBER SPANGLER: The measure

application itself that you submitted.

MS. VICALE: It should be on the worksheet.

MEMBER SPANGLER: The worksheet.

MS. GORDON: Same document we had emailed you last week including all the preliminary analysis and comments.

MR. SPEAKER: Is somebody able to share that document?

MEMBER SPANGLER: I'm assuming it's the results from the 17 entities that did pass validation, because that's the sentence that's right before that, but I just wanted to clarify.

MS. GORDON: Right. And that sounds like that would -- sorry you guys. Right. And that would be -- that's correct because where it talks about two of the 19 entities, that was what was discovered during the validation process and, therefore, only 17 of the entities actually publicly reported at that time, if I'm understanding what you were asking.

MEMBER MARRS: To follow up on the stain

piece, you kind of clarified why stain intolerance can be included, but you list that antiplatelet contraindications are included. And can you describe how that's defined?

MS. GORDON: Yes. The reason that we felt like we could include those contraindications were that there were certain diagnoses that seemed to apply across the board for that. Such things such as an intercranial bleed, a GI bleed.

And so, basically we are using more of ICD-9 coded type diagnoses to identify that. And so far we weren't able to find the same, you know, a parallel with the statin use, which is why we would really like to try to work with -- at least it's our member's goal, you know, to try to find a way to get a discreet feel, but can identify like a flag of some sort that this patient is allergic to a statin so that we knew that it was accurately identifying that that's what they are having a contraindication to.

DR. KAUFMAN: The coding specificity for statin intolerance is just not specific enough to really pick up through claims data.

MEMBER MARRS: Right. It was more along the lines of if you were pulling any drug allergy data in to identify that or if it was ICD-9 based.

CO-CHAIR GEORGE: Tom James, did you have a comment?

MEMBER JAMES: No, I don't think I had my hand up.

CO-CHAIR GEORGE: Jason.

MEMBER SPANGLER: Thanks, Mary. I just want to follow up. So, like Sana said, I agree. I thought the outline was really good about how you're going to do that.

And the only numbers that I see are here, but I'm not even sure how to interpret these numbers. It's just a range. Is there a median? Is there -- so, is there any kind of actual data on what the validity, I mean, is this for the 17 that you say passed?

I mean, so was 44.8 percent at the group level, and 37.92 percent, that was considered pass?

MS. GORDON: No, not exactly. It would

-- what we do is we look at patient-level data and then work with our members on how this data is being obtained out of their EMRs.

And so, in -- I guess basically in not passing, there were just too many instances of where when we looked at patient-level data it wasn't -- it just didn't -- I don't know how I'm trying to say this. It just didn't really pass.

There were too many questions about the data that they were submitting to us through their EMR.

MEMBER SPANGLER: So, I guess my question is, what does it mean that these 17 entities passed? What does that mean that they passed? Do they have a certain -- what does that mean?

MS. GORDON: Right. Right. When a random sample of patients were looked at that they met that -- with the data that was in their EMR, that they passed that --

MEMBER AL-KHATIB: I think what he was trying to get at -- what he's trying to get at is

what percentage of agreement did you consider passing? Percentage.

Like, do they have to meet 90 percent like agreement between what's in the EMR and what you captured through the billing codes, or what is that?

MS. GORDON: I know, and I don't have that information with me. Is it something that we can --

MEMBER AL-KHATIB: I mean, the question for me is -- because it looks like they have a very good plan for validity, but I think we need more quantitative --

MS. GORDON: Sure.

MEMBER AL-KHATIB: -- data to tell us exactly what they found. Now, the plan, the plan makes perfect sense. So, I wonder if we could defer until we get this information from the developer.

MS. WILBON: Yeah. So, that might actually be somewhat soon just because we're losing another member, which I believe will put us below quorum. So, we would actually, I mean, we could

continue to discuss, but we couldn't, you know, pass any votes at this point with the committee any more measures without, you know, having to do a lot of work to get everyone back on the same page.

It probably makes more sense if everyone in the room is okay with that to pause this discussion and we can kind of take some of the concerns that the committee has, follow up with the developers.

We have two follow-up calls scheduled. We'll make sure you guys have that information and we'll kind of get everyone ready to continue the discussion.

And maybe by that time you guys will have an opportunity to provide the committee with some additional information that might help clarify some of these issues.

So, I think that might actually be a good point for us to pause.

Does that sound okay, Mary?

CO-CHAIR GEORGE: Yeah, I think so, because I think you know kind of a little bit more

about what the committee is asking in terms of those validity testing results, but we can't vote on it.

MS. VICALE: So, we will go to member and public comment now for the measures reviewed on Day 2. This is roughly 20 minutes early from when the original time for a public comment is scheduled for. Only 3:45. So, we'd like to just note that.

So, also of note there will be member and public comment for all of the measures. And that will be from October 16th to November 16th.

So, I'd like to ask the operator to open up the line for member and public comment right now. And please keep that line open a little bit longer just in case folks would like to comment knowing that this is a little bit earlier than originally scheduled.

MS. WILBON: Also, if there's anyone in the room while we're waiting for the operator to queue up if you have any questions, please feel free to step up to the microphone or -- okay. Doesn't look like there's anyone in the room.

MS. VICALE: Operator, please open the

line for member and public comment.

THE OPERATOR: If you'd like to make a public comment at this time, please press *1 on your telephone keypad.

(Pause.)

MS. VICALE: We have a public commenter in the room.

MS. SLATTERY: So, hi. Lara Slattery from -- it's green. Yeah, it is green. From the American College of Cardiology.

I was just wondering if you could clarify the earlier vote today regarding -- I lost my measure. Yeah, Measure 0070. Just so we could understand what potentially could happen with the process, because I'm still a little unclear.

Was the vote that occurred today specific to the e-Measure, or does that actually revisit the endorsement status of the existing endorsed measure?

MEMBER CHO: It's my understanding from Helen when we started this discussion, it's only for e-Measure and that we would be voting on the paper

measure or the registry measure at another time.

MS. WILBON: That's correct. We'll review all of the registry versions of those measures. I believe there's three measures on our follow-up call that's scheduled.

So, today the committee voted on the e-Measure specs.

MS. SLATTERY: And is that happening within this phase?

MS. WILBON: Yes, it will be like in two weeks. A week or two weeks.

MS. SLATTERY: Okay. Thank you.

MS. VICALE: Is there anyone else on the line that would like to make a public comment, or anyone else in the room?

(No comments.)

MS. VICALE: And to note for the record, there were no comments or questions posed through the chat window in the web platform.

THE OPERATOR: And there are no public comments.

MS. VICALE: Okay. Well, thank you

very much. Again, I'd like to reiterate that the draft report will be posted for public comment. And that will happen from October 16th to November 16th.

CO-CHAIR GEORGE: I'd just like to thank everyone. I know this is a really difficult task that we have. And thank you for your good comments and discussion today.

MS. VICALE: Before we leave for today and adjourn the meeting, I'd like to just go over the timeline and the next steps for the project.

As you can see here, the post-meeting call will be held September 25th, from 2:00 to 5:00 p.m. eastern time.

We scheduled a second post-meeting call for October 9th from 2:00 to 4:00 p.m. as well. And as you can see as already noted, the draft report will be posted for public comment October 16th to November 16th. And we will have a standing committee call to review any comments received during that time on December 7th from 1:00 to 2:00 p.m. eastern time.

A draft report will be posted for NQF member vote from December 18th through January 5th -- I'm sorry, that should say January 5th, 2016.

And the CSAC will review the recommendations from the Standing Committee and provide their recommendations to the Board on January 12th of 2016.

And endorsement via the Board will happen during February 2016. Exact date we will follow up with as the project continues. And the appeals period is from February 8th to March 8th of 2016.

So, on behalf of the NQF staff I'd like to thank the committee for their hard work throughout the past two days.

I'd also like to thank the developers for presenting their measures. And I'd like to thank the public for joining us for our in-person meeting to evaluate the measures.

So, thank you all very much and we look forward to having you on our post-meeting call.

(Whereupon, at 3:28 p.m. the meeting was

adjourned.)