NATIONAL QUALITY FORUM

+ + + + +

SCIENTIFIC METHODS PANEL

+ + + + +

MONDAY

OCTOBER 28, 2019

+ + + + +

The Scientific Methods Panel met at the National Quality Forum, 5th Floor Conference Room, 1099 14th Street, N.W., Washington, D.C., at 9:00 a.m., Dave Cella and David Nerenz, Co-Chairs, presiding.

PRESENT:
DAVID CELLA, PhD, Northwestern University; Co-Chair
DAVID NERENZ, PhD, Center for Health Policy and Health Services Research, Henry Ford Health System; Co-Chair
J. MATT AUSTIN, PhD, Armstrong Institute for Patient Safety and Quality at Johns Hopkins Medicine
BIJAN BORAH, MSc, PhD, Mayo Clinic
JOHN BOTT, MBA, MSSW, Healthcare Ratings, Consumer Reports
DANIEL DEUTSCHER, PT, PhD, Maccabi Healthcare Services
LACY FABIAN, PhD, The MITRE Corporation
MARYBETH FARQUHAR, PhD, MSN, RN, American Urological Association
JEFFREY GEPPERT, EdM, JD, Battelle Memorial Institute
LAURENT GLANCE, MD, University of Rochester School of Medicine and Dentistry
JOSEPH HYDER, MD, Mayo Clinic
SHERRIE KAPLAN, PhD, MPH, UC Irvine School of Medicine
JOSEPH KUNISCH, PhD, RN-BC, CPHQ, Memorial Hermann Health System
ZHENQIU LIN, PhD, Yale-New Haven Hospital
JACK NEEDLEMAN, PhD, UCLA
EUGENE NUCCIO, PhD, University of Colorado, Anschutz Medical Campus
SEAN O'BRIEN, PhD, Duke University Medical Center
PATRICK ROMANO, MD, MPH, UC Davis
SAM SIMON, PhD, Mathematica Policy Research
ALEX SOX-HARRIS, PhD, MS, Stanford University
CHRISTIE TEIGLAND, PhD, Avalere Health
RONALD WALTERS, MD, MBA, MHA, MS, University of Texas MD Anderson Cancer Center
TERRI WARHOLAK, PhD, RPh, CPHQ, FAPha, University of Arizona, College of Pharmacy
ERIC WEINHANDL, PhD, MS, Fresenius Medical Care North America*

NQF STAFF:
MICHAEL ABRAMS, Senior Director
SHANTANU AGRAWAL, MD, MPhil, President and CEO
KAREN JOHNSON, Senior Director
ANDREW LYZENGA, Senior Director
ELISA MUNTHALI, Senior Vice President
YETUNDE OGUNGBEMI, Project Manager
SAM STOLPE, Senior Director
ASHLIE WILBON, Senior Director

ALSO PRESENT:
MARK ANTMAN, STS*
MARY BARTON, NCQA
ANNA CHRISTENSEN, Mathematica*
CINDY CULLEN, Mathematica*
BOB REHM, NCQA
DAN ROMAN, NCQA
DAVE SHAHIAN, STS*

*** present by teleconference**

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

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

9:06 a.m.

MS. JOHNSON: Welcome to our Fall 2019 cycle Scientific Methods Panel meeting. This is pretty exciting for us. We've been able to gather people together in the past for some methods discussions.

Today we will be doing measure reviews. So we've done these over the phone before, I think we'll probably -- It will probably be more fun, more interesting, more interactive being able to do it in person.

It's always good to be able to see body language, get to know your fellow panelists. So thank you for coming. I am Karen Johnson. I am one of the senior directors here at NQF and I have the pleasure of getting to work with the methods panel.

So you guys take up a lot of my time and thank you so much --

(Laughter.)

MS. JOHNSON: So I hope you guys enjoy

this interaction. We are going to be doing -- We are going to start off with our usual, if we could just go to the next slide, we'll do our welcome introductions and disclosures of interest. We do have to do those today.

Then we will spend a little time going over process and just a few reminders about our criteria and then we will get into our measure evaluations.

And I'll give you -- Where should I be pointing? Is it --

MS. WILBON: Nowhere. Just anywhere.

MS. JOHNSON: Anywhere.

(Laughter.)

MS. JOHNSON: Anywhere isn't working. There we go, okay, anywhere works, okay. So content leads, we're going to just do a real quick introduction in a couple of minutes here of our staff, but before we do that we will have Dave and Dave say hello, our esteemed co-chairs, and then we'll go around and introduce our staff, including Shantanu, who, sorry, we didn't put you

on our slide.

MR. AGRAWAL: Don't worry.

MS. JOHNSON: But Shantanu, our fearless leader. Dave and Dave, would you like to welcome folks to the meeting.

CO-CHAIR NERENZ: Sure. Welcome and we're looking forward to a good couple of days. I guess the first thing we should do, we tried a couple of years ago to establish a Dave and David distinction and we couldn't hold it. I couldn't even hold it myself.

(Laughter.)

CO-CHAIR NERENZ: So I think we are Dave N. and Dave C., at least that is a reliable distinction. Otherwise, we can't remember which is which.

It's great to have you here. I've always enjoyed the face-to-face meetings. It's great to see people and have the discussion. I always learn a lot from these and I think this is going to be unique.

As everybody knows it's the first time

we've done this particular approach to discussion on measures that did not retrieve consensus in the subgroup.

So I've lost a little sleep over this because I am thinking that it's probably as easy to get a disagreement among 25 people as it is among six people and then I'm not sure where we go.

But it will be interesting learning along the way, so I am glad you are here and glad you took time to do this and I'm looking forward to a good couple days.

CO-CHAIR CELLA: Hi, everyone. I am Dave Cella, just add my welcome and thank you for all the time and attention that you are paying to these important reviews.

Do you want to say something about the consistency now or later?

(Simultaneous speaking.)

MEMBER ROMANO: We're having trouble hearing at this end.

(Simultaneous speaking.)

CO-CHAIR CELLA: Okay. Hi. I'm Dave Cella and I'll just talk a little bit louder and add my welcome to Dave N.'s welcome.

MEMBER ROMANO: Are your microphones in the ceiling or --

MS. WILBON: Yes, the green discs. The green discs.

MEMBER ROMANO: Oh, the green discs, okay.

MS. JOHNSON: Yes, so what we may do is you may still have to use your outside voices in here. So we will let you know if we can't hear you on this end and you guys just let us know if on that end you can't hear us.

MS. WILBON: And it's being recorded?

MS. JOHNSON: It is being recorded.

MS. WILBON: So you can hear everything we are saying?

MS. JOHNSON: Yes. These will go -- When we go on breaks you'll notice that they will turn red. That means they are muted. If they are green we are not muted.

MEMBER ROMANO: Is there a way to turn up the amplification or is that not an option?

MS. MUNTHALI: So we do have some mics. There are about four mics, we'll pass those around. We did this last week with the CSAC.

We apologize, we just moved in and so this is part of being in a new home, but we did have our IT colleagues working on the sound, so they are going to come to make sure that everything is okay for us.

MEMBER KUNISCH: It's a beautiful space.

MS. MUNTHALI: Thank you.

MEMBER KUNISCH: It's visually very nice.

MS. MUNTHALI: But I think the folks on the phone can hear us perfectly, which is very ironic.

(Laughter.)

MS. JOHNSON: All right. Let's go around this way and we'll get staff and then

we'll get staff on this side and then we'll get Elisa to go over the DOIs.

MS. MUNTHALI: Okay. So hello, everyone. My name is Elisa Munthali. I am the Senior Vice President for Quality Measurement and I wanted to thank you so much for all of the time you put into the Scientific Methods Panel and for being here in person and for those on the phone thank you for joining us virtually.

I know a couple couldn't make it, but you are all so committed and to give you a lot of work and we so appreciate the volunteer hours that you put into this, so thank you. Shantanu?

MR. AGRAWAL: Yes. Yes, I was working on the sound. So I am Shantanu Agrawal, the CEO of NQF. Thank you all for being here. I really appreciate all the time that you take to work on this extremely critical and important committee.

I do think it would be highly remiss of us, there is a lot of intellectual brainpower in the room, lots of statistical and methodological power.

I do want the over/under on a Game 7 of the World Series. It's going to be really important for the entire committee and for the entire city as it turns out. So, again, thank you for being here.

MR. LYZENGA: Hi. Andrew Lyzenga, Senior Director here at NQF. I am looking forward to getting started for the day.

MR. ABRAMS: Yes, good morning, everybody. I am Michael Abrams, also a Senior Director here at NQF, and excited to be working on this effort with you all.

MS. WILBON: Good morning, everyone. My name is Ashlie Wilbon. I think it got a chance to say good morning to everyone, but there may be a few that I missed, so, welcome.

We are excited to have you guys here today and, yeah, I hope everyone has fun today.

MR. STOLPE: Good morning, everyone. Sam Stolpe, also a Senior Director here at NQF, and I just wanted to add a welcome. I didn't get to say hello to everyone, but I especially wanted

to welcome the newcomers to the panel. Thanks for making the trip out.

MS. OGUNGBEMI: Hi. I am Yetunde Ogungbemi. Good morning, everyone. I will reiterate Karen's comments, you take up a lot of my time, but it's a pleasure. Welcome, thank you.

MS. JOHNSON: Okay. And with that I think we'll just move straight into -- Well, in just a second, before that I want to make sure that we congratulate our friend Dave Cella on his election to the National Academy of Medicine.

(Applause.)

CO-CHAIR CELLA: Thank you.

MS. JOHNSON: Speech?

(Laughter.)

CO-CHAIR CELLA: I, you know, it's just one of things where, you know, you sort of generally say why me. I mean it's really quite an honor and I was pretty surprised, so thank you.

MS. JOHNSON: Congratulations.

CO-CHAIR CELLA: Thank you.

MS. JOHNSON: Yes. All right, Elisa, do you want to walk us through DOIs?

MS. MUNTHALI: Yes. So welcome and hello, everyone, again. So when you were first selected to be a part of this committee we sent you a pretty lengthy disclosure of interest form and information and what we are asking you today is to orally disclose what you gave us in written form.

We do not want you to recite your very impressive resumes, but there are a couple of things in particular that we are looking for. We are interested in any grants or research or consulting that is related to the work of the Scientific Methods Panel.

And before we go around the room, so what we will do is go around the room clockwise to my left and then to I think Jen who is on the phone and Eric who is also on the phone.

And so what I would like you to remember is that you sit on this group as an

individual. You do not represent anyone who may have nominated you for this committee or your employer.

We are interested in both paid and unpaid activities but only as they are related to the work in front of you. And I think this is the most important reminder, just because you disclose does not mean you have a conflict of interest.

We do this in the interest of transparency and openness. So I think we'll start with Christie, yes.

MEMBER TEIGLAND: And I have nothing to disclose.

MS. MUNTHALI: Perfect. Thank you. And if you can just introduce yourself and let us know who you are with.

MEMBER TEIGLAND: Oh, sure, yes. Hi. Christie Teigland. I am a Principal of Health Economics Advanced Analytics at Avalere Health. They were acquired by Inovalon about three, four years ago now almost, and I was previously with

Inovalon, so I do a lot of big data research.

MEMBER NUCCIO: Gene Nuccio, University of Colorado Anschutz Medical Campus. My background is primarily in home health, post-acute care, and so I've worked on the Whole Health Measures, the value-based purchasing, and some MedPAC work, but I have no conflict with any measure today.

MEMBER BOTT: John Bott. I am an independent contractor. Currently I am doing some work helping Thomas Jefferson University Employer Health Care Alliance Cooperative in Wisconsin, the Leapfrog group, a couple attempts, I don't know if it's salient or not, selected for the AHRQ ESIs, which essentially -- manages that contract, and also I am a CMS Hospital Star Ratings -- that's all.

MEMBER WARHOLAK: Good morning. I am Terri Warholak and I am a professor at the University of Arizona College of Pharmacy.

I have been working on psychometrics for many years and teach that and I have also

been working with the Pharmacy Quality Alliance as a volunteer for many years as well.

I have no conflicts with any of the measures today.

MEMBER FARQUHAR: Good morning. My name is Marybeth Farquhar. I am the Executive Vice President for the American Neurological Association.

I am six months new into that job. I am in charge of guidelines, measure development, and research as well as the data registry that we are hosting, and I have no conflict of interest.

MEMBER DEUTSCHER: Hello, good morning. My name is Daniel Deutscher. I work out of Israel. Maccabi Healthcare Systems is a public health plan in Israel where I serve as the National Director of Research.

I have been involved in measure development, patient-reported outcome measures for many years and also involved very much in creating large databases, big data analysis using patient-reported outcome, lots of treatment data,

patient data.

I am a -- I also serve as a consultant for FOTO, Focus on Therapeutic Outcomes, they are submitting to this committee measures on a regular basis, so this is the disclosure I have.

MEMBER HYDER: Hi. My name is Joe Hyder. I am anesthesiologist IC physician and epidemiologist at Mayo Clinic in Rochester, Minnesota.

I have worked for and have some work with pro bono the American Society of Anesthesiologists, their committee on performance and outcome measures and core measures -- but I have no conflict.

MEMBER SOX-HARRIS: I am Alex Sox-Harris. I am a health services researcher at the Department of Veteran Affairs in Palo Alto and at Stanford University Department of Surgery.

I have federally funded research looking at the measurement properties of various quality measures. I think there is one measure today, it's bereavement that is put forward by

the VA and I know some of those people, so that is a potential conflict.

I have worked with some professional societies helping develop quality measures, but nothing that is in front of the group today.

MEMBER BORAH: Hi, everyone. I am Bijan Borah. I am a professor of health services research at Mayo Clinic and I have nothing to disclose with regard to these measures.

MEMBER LIN: Hi. My name is Zhenqiu Lin. I am a Senior Research Scientist from Yale School of Medicine and Senior Director of Data Management and Analytic at Yale CORE.

So I have been working on measure development for, oh, over a decade, and we will have some measures before this panel.

MEMBER O'BRIEN: Hi. I am Sean O'Brien. I am a biostatistician at Duke University. I have been working on measure development for probably about that long or a little longer.

I previously served on a handful of

NQF panels and task forces, including the measure testing task force and the panel for risk adjustment and socioeconomics and status and I was a statistical consultant to NQF a couple times.

I work with national sample registries from the Society of Thoracic Surgeons and American College of Cardiology, so I do have conflicts with a handful of the measures, or a couple of the measures that we are doing this weekend.

MEMBER AUSTIN: Yes, good morning. I am Matt Austin. I am on the faculty at the Johns Hopkins Armstrong Institute for Patient Safety and Quality.

Probably my most significant contract is to provide guidance around measurement equities and there is no conflicts.

MEMBER GEPPERT: Jeffrey Geppert, Battelle Memorial Institute. My only disclosure, brief work on contracts with the Centers for Medicare and Medicaid Services to manage their

portfolio measures for accountability programs, including NQF endorsement.

MEMBER ROMANO: I am Patrick Romano. I am a general internist and general pediatrician. I am employed by the University of California Davis Health, UC Davis Health, in Sacramento, California.

I have worked on measure development and testing for 25 years, or more. Currently my group has contracts with AHRQ related to the hospital quality indicators and CMS related to patient safety indicators.

We also work with UCSF on a contract to develop a new quality measure around a CT dose, radiation dose, which is funded by CMS. So through these relationships I work with Impact International on what is called the CMS Patient Safety Measures contract and our team has one measure that is before this committee now, which is 3533(b), I think.

MEMBER SIMON: All right. Sam Simon. I am a Director of Quality of Measures at

Mathematica. We are a CMS measure developer. We are a contractor for the Centers for Medicare & Medicaid Services and one of the measures that I advised is up for committees the measure --

MEMBER KUNISCH: Good morning. Joe Kunisch. I am with Memorial Hermann Health System in Houston. Go 'stros.

(Laughter.)

MEMBER KUNISCH: We'll talk about that.

(Laughter.)

MEMBER KUNISCH: But I oversee all their regulatory reporting programs and our organization also participates in multiple electronic clinical quality measures as seen with the CMS contractors and I do not believe any of them are up for review. That's all.

MEMBER NEEDLEMAN: I am Jack Needleman. I am a Professor and Chair of the Department of Health Policy and Management at the UCLA Fielding School of Public Health.

I am a health economist by training.

I have previously served and continue to serve, previously served NQF as a member of the Technical Expert Panel on the hospital-related nursing quality performance measures.

I currently serve, and have served since the beginning, on the Cost and Resource Use Committee. I have provided some unpaid consulting to the American Nurses Association as they have tried to move the nurse staffing measures through the NQF endorsement process and have no current conflicts of interest with any of the measures today and as far as I know no conflicts with any of the other members of the committee.

(Laughter.)

MEMBER KAPLAN: I am Sherrie Kaplan and I psychometrician by training. I am a Professor of Medicine and Assistant Vice Chancellor for the Healthcare Measurement and Evaluation at UC Irvine.

I think the UC's are represented in this room. But, anyway, so I am on the NQF

Patient Experience and Function Committee and I have, currently my husband and I are unpaid consultants to the Danish Government for a planning initiative on diabetes quality for their country.

I don't think that puts me in conflict with the measures because I don't really have a vested measure developer role in that. I have grants under review and consideration from AHRQ for the child health rating inventories. It's a measure of animated measure of children's self-reported health ages 4 to 12.

MEMBER WALTERS: My name is Ron Walters. I am a medical oncologist at MD Anderson in Houston. I have nothing to disclose and I'm, same thing, the only conflict I had with the prostate measure was I suggested we needed one. I didn't have anything to do with development.

By the way, for those of you that watched the game, there are only two times, 1906, 1996, they showed this on a little graphic, that

the home team has not won a game during Game 5.

They didn't say what the statistics were for the sixth game, but you can rest with some hope that if that holds true you're in good shape.

(Simultaneous speaking.)

MEMBER GLANCE: Good morning, everybody. Hi, my name is Larry Glance. I am from the University of Rochester. I am a cardiac anesthesiologist and a health outcomes researcher. I am professor and vice chair for research in my department.

I don't believe I have any conflicts of interest. My disclosure is that I also serve on the American Society of Anesthesiologists Committee for Performance and Outcomes Measurement and we do develop quality metrics for anesthesiologists, none of which I believe are in front of this committee.

MEMBER FABIAN: Good morning. My name is Lacey Fabian. I am a research psychologist by training currently at the MITRE Corporation. My

background is always focused on quality measures, particularly from the development side and then from the best practices side.

My work is also now shifting to focus more on clinical decision support as well.

CO-CHAIR CELLA: I am Dave Cella from Northwestern University. I chair the department called Medical Social Sciences. I have worked as a consultant and a PEP member for Yale, RAND, and RTI. Yale and RAND current, RTI in the past, but none of those, there are no measures in front of the committee for this meeting that are related to anything I have advised on.

CO-CHAIR NERENZ: Dave Nerenz. I am Director Emeritus now of the Center for Health Policy, Health Services Research at Henry Ford Health System in Detroit.

I have worked at the intersection of quality measurement and racial and ethnic disparities of care for over 25 years. I was co-chair of the NQF panel on socioeconomic risk adjustment.

I have published a few things about quality measures and their effect on providers who care for vulnerable populations. So I guess I have some opinions out there, but I don't have conflicts of interest.

I am not a measure developer. I don't have a connection to anything in front of us these two days.

MS. MUNTHALI: Thank you, everyone. Oh, sorry.

CO-CHAIR CELLA: May I just ask, because there are some new people here and it's a long table I just want to call your attention to the role models of Sherrie Kaplan and Eugene Nuccio.

If you could tip your -- Thank you. Tip your name cards in 30 degrees so we can see them from here that would be very much appreciated. Thank you.

MS. MUNTHALI: Thank you for that reminder. And so we'll go to the phone, Jennifer Perloff. Jen, are you with us perhaps on mute?

(No audible response.)

MS. MUNTHALI: Okay. Eric Weinhandl?

MEMBER WEINHANDL: Yes, good morning.

MS. MUNTHALI: Good morning.

MEMBER WEINHANDL: This is Eric. Sorry I am unable to join you this time, a conflict with one of my children's schedules.

So I am an epidemiologist biostatistician. I am currently employed by Fresenius Medical Care North America. It's (telephonic interference) health services provider as well a manufacturer (telephonic interference) products.

I am an epidemiologist biostatistician myself. I have worked in chronic kidney disease and dialysis specifically since I graduated from college, all things really.

So any time that there is a metric that comes up that is related to that in some way, shape, or form I probably have some kind of convoluted conflict of interest.

In the case of today's measures before

the committee I believe the (telephonic interference) issue is the one that would be more directly pertinent to my conflicts. I (telephonic interference) topic is the specific case of this metric I have focused on a pro bono basis to the Care Quality Alliance, which is the measure -- it's good to be here.

MS. MUNTHALI: Thank you so much, Eric, and thank you for being with us virtually. Before I turn the meeting over to Ashlie and Karen I just want to remind you that if at any time you remember that you have a conflict we would like you to speak up.

You can do so in real time or you can come to any one of us on the NQF staff. Likewise, if you believe that one of your colleagues was acting in a biased manner we want you to speak up.

So thank you very much and I will turn it over to Karen.

MS. JOHNSON: So thank you, everybody, and in terms of recusals we will talk about that

a little bit later in the morning, but we have -- I am also going to go through our meeting materials that when we get to that part you'll notice that we have noted who is recused on certain measures.

I am going to ask you guys to help us be honest on this. If we missed somebody please make sure that you let us know that we missed and that you should be recused from a measure, but we'll get into that as we go.

So we wanted to do a little bit of a, just a reminder, perhaps a little bit of a reminder for some of you and maybe a little bit of a reminder for people who are a little bit new to the process.

So first of all our Scientific Methods Panel was formed in 2017, so you guys are two years old now. You are one of our newest groups here at NQF and the formation of the Scientific Methods Panel was actually a direct outcome of the Kaizen Event that we did here in the Spring of 2017 really to try to help us not only become

more agile in our endorsement process, but also hopefully to become more consistent in the process and to take some burden off of our already empaneled standing committees.

So you have a two-fold charge. One that you had intimate experience with in the last several weeks, and that's conducting the evaluation of conflicts measures for reliability of validity, and then also you serve as our advisors on methods, on our criteria, on, you know, what's new in the measurement science world, so we look to you for that advice.

So the evaluation process, this may be a little bit new to you, maybe not to everybody, but what happens is developers, or stewards, submit specifications and the testing attachment to us at certain times in the year, and that's called our intent to submit deadline.

And we take that information when it comes in the door and we determine which measures are sent to the methods panel. So we don't send you every measure that comes in for potential

endorsement, we send you the ones that we call complex measures.

The complex measures are outcome measures, intermediate health outcomes for patient-reported outcomes, any cost measures that come in the door or resources measures, any composite measures, and any instrument-based measures, okay.

So those are the complex ones. Those are usually the hard ones, right, that you guys get. So we figure out which ones those are and send them to you.

And while we are doing that we, even before we send them to you we do try to take a first look and just make sure that some of our minimum requirements are met before we send them out to you.

So we want to make sure that testing is done at the levels of analysis and that sort of thing that they have been specified for. And as you know different measure types we have different requirements for levels of testing that

we require.

In the summer, this summer actually -- Actually, no, not this summer, this year, so calendar year 2019 we said it is time that we really start expecting to see testing done with ICD-10 data as opposed to ICD-9. So we've had a several year grace period, but we are really trying to invoke the ICD-10 testing.

We also want to make sure that things that need to be answered in our submission forms actually have a response and we also allow, well we don't allow, we expect by the time a measure comes back around for maintenance we expect empirical testing for validity.

Now that is not always going to be possible for every measure developer. So if they are still relying on face validity even at the maintenance evaluation we ask for a justification of why they weren't able to do empirical testing, so we make sure that at least that is included in there.

All right. Then we ship it out to

you. No, before we do that, let's see. There is a lot of steps in here that I am sure you guys are aware of.

We ask you guys to submit your measure-specific disclosures of interest. So this is kind of the overall kind of public one, but we ask for the measure-specific ones and that's what tells us who needs to be recused from the discussion.

We then assign measures to subgroups. So the number of subgroups depends on the number of measures that have come in the door that are going to you guys and who gets which measures depends on your expertise, need for recusal, previous experience, I'll explain that in a minute, and then, obviously, the number of measures, so the more measures we have the more that you probably will be assigned.

In terms of previous experience, we do note, you know, sometimes measures have come in and they haven't made it past the methods panel, so we try, if that's the case, we try to make

sure that at least one or two people who have reviewed a measure before has a chance to review it again.

It can't be, just the way it works out, it can't be exactly the same people, but we do try to have some overlap there. We give you four weeks to complete those initial evaluations and as you know you fill out that form that we have, we call that the preliminary analysis for, or here at NQF we will call it the PA, because we love using acronyms, so the PA form.

And then finally when you submit those to us we collate those evaluations and ratings, we figure out did a measure pass, did it not pass, did it fall kind of in between pass and no pass, and we call that consensus not reached, or CNR, okay.

So that, and you can see that in your discussion guide kind of where things landed on the ratings, okay.

Now this fall we actually have some new process and you're seeing part of that today

because you are here in D.C. with us in a big room as opposed to doing all of this over the phone over like eight different calls in a 2-week period.

So we are allowing, let's see -- Do I have this in the right order? I hope I do. We have given you guys opportunity to pull measures for discussion.

So once you see where those ratings are you can say, hey, this measure passed, but I'd still like to talk about it, so we gave you that option to do that. If you did that then that is on our agenda for today and tomorrow.

We have developed a discussion guide and figured out the agenda, so our ordering, and that is based on those initial ratings and EPAs. It's based on whether or not developers have provided additional information and whether any of the measures were pulled, and we also as staff have the opportunity to pull measures for discussion.

I missed one really important change

to our process. I don't know how I missed it, but we were able, because we were able to have this in-person meeting and we pushed this meeting a little further in a timeline than we were able to push those phone calls that we had before, that allowed us a small amount of time, not a huge amount, but some time to actually let developers respond to your initial evaluations, okay.

So we took those evaluations, we combined those, we sent them out to developers, and developers could choose, they didn't have to, but they could choose to make responses in written form to you guys.

Most of them actually did take us up on that and provided some written responses and hopefully we are able to speak to the main concerns or even all of the concerns that you raised in your evaluations.

All of those things we at NQF are making available to the public. That's also a change from what we did before. So before we had

these discussion guides and those were pretty much we gave those to the developers, we gave them to you, but nobody else saw them.

Now everybody has these. So these are posted. If they are not posted, they will be. I think they are already up on our website, so members of the public who are following along you should be able to see those.

Obviously, we have our in-person meeting and we will vote today and first thing tomorrow on some of these measures and then, finally, we will provide summaries of your discussion and votes to our standing committees.

So that's changed a little bit from what we have done before. Before we did not do that if the measure did not pass the methods panel.

Now even if the methods panel does not pass a particular measure we will still provide that detailed summary to the standing committees and the standing committees will have an opportunity to discuss those measures should they

desire to do that.

Let me stop there and see if anybody has any questions about our new process. And, Sherrie, thank you for doing that. I have -- I meant to give you a few housekeeping -- Sherrie has been here before.

Just a couple of reminders before we go in, this meeting is open to the public as we talked about, so people can be listening on the phone.

It is being recorded and the fellow down at the very end with the headphones on is our court reporter, so he is making sure that we have a really good transcript of the meeting today and tomorrow.

For those of you on the phone you are already doing a great job, but we would ask that you would mute your lines when you are not speaking and if you do want to speak be sure to tell us your name so that we know who you are, and, SMP members, methods panel members who are on the phone you can certainly use our chat

function in our web platform and let us know if you want to talk, but you should also just feel free to jump in.

You are at a little bit of a disadvantage, so just, you know, if you need to talk just talk. And, finally, well two more things, your tent cards, if you guys in the room want to talk we ask that you put your tent cards up like this and that signals us that you want to say something.

And then, finally, when the time comes we will have developers and/or stewards in the room to help answer questions. If they aren't in the room we will actually allow them and ask them to sit at our table, so that's why we have a couple of chairs over here set aside for our developer colleagues. So now, Sherrie.

MEMBER KAPLAN: I just wanted to remark on and ask about the communication between the standing committees and this panel, because there is, as having sat on one of them there is some confusion about, you know, well you guys

didn't disagree.

Well, you guys, but the idea that, you know, their communication, our communications with them is right now I don't think optimized and I think, you know, a better understanding by the committees might help, especially for one struggling as the patient experience committee is with this patient level error time we were talking about averaging across patients, within patient across items and then across patient and in between providers.

There is a lot more of that that goes on in that committee. So our communication with them I don't think right now is as clear as it could be.

MS. JOHNSON: It's a good question and so far we haven't figured out how to make that, so let's put that on our parking lot for tomorrow and let's talk about how, you know, ideas that we might have to do that.

Some of you guys, as you mentioned, are on these panels so you can be kind of the

liaison, if you will, to those standing committees, but that's not true across the board and what we have done is we have expected staff to take on that kind of liaison role and that may or may not have worked out as well as we had hoped.

We were really -- There is a lot of meetings of our standing committees and it didn't feel fair to ask you guys to do more so we haven't asked you guys to actually go to our meetings, but maybe that's something we want to at least talk about.

We just kind of decided that we wouldn't do that to you.

MEMBER KAPLAN: Yes. Well I was just going for a little more clarity on the role of this committee to help us understand exactly what we are doing for you.

MS. JOHNSON: Okay. Let's --

MEMBER KAPLAN: And you don't have to do that now.

MS. JOHNSON: Yes, let's think about

that tomorrow for sure because we as staff actually write those descriptions of what happened.

Maybe we need to have you guys look at those and make sure we're getting it right. That might help. We think we are, of course, we don't write out stuff that we think is wrong, right, but we might not be clear enough so maybe that's something we can do as well. Thank you. Jack?

MEMBER NEEDLEMAN: Jack Needleman. So this is less of -- This is not a question, it is a comment and really directed to some of the new members, so, of course, they are going to be coming into some conversations that have been underway, so you have not talked about some of things that the switch to this committee has produced in terms of the review process.

I have seen two things that I think are particularly relevant and the developers are still adjusting to it and we're still trying to sort out our standards.

One is because there has been a lot

more methodological expertise looking at the measures the standards I believe for reliability and validity have been ratcheted up.

Now others on the committee may not agree about that, but my sense is we're applying higher standards than have historically been applied and the developers are adjusting to that.

The second thing is as we saw in many of the submissions the developers are using the standards that are sort of out in the literature for what's a strong relationship and an acceptable level of conformance, whatever the measure is.

And as we have looked at some of the data related to what that means in terms of ranking and stability of ranking there has been a lot of discussion on the committee about whether those standards are frankly rigorous enough.

We have not resolved what our new standards are, but that keeps coming up in conversation among us and I expect will come up again today.

MS. JOHNSON: Thanks, Jack. Patrick, go ahead.

MEMBER ROMANO: Yes. Patrick Romano. Could you clarify when you say that the SMP meets in person to discuss and vote?

So is the entire SMP voting on every measure or are people only voting on the subcommittee that they served on?

MS. JOHNSON: The answer is yes to both. So I will actually get into that in a few minutes and I'll explain how we're going to have you guys vote.

MS. WILBON: Karen, should we tell people where bathrooms are and stuff like that?

MS. JOHNSON: Oh, yes.

MS. WILBON: I actually don't know where the bathroom is.

MS. JOHNSON: Ashlie suggested we let you know where the men's room and ladies' rooms are, and that's actually a good question, where are they? They are --

(Laughter.)

(Simultaneous speaking.)

MS. MUNTHALI: They are just the reception desk and then to your immediate left is the women's restroom and the men's restroom is just down the hall. I guess Andrew has confirmed that.

MR. LYZENGA: On the other side.

MS. MUNTHALI: Yes, on the other side.

MS. JOHNSON: Okay. Is that clear as mud?

MS. MUNTHALI: On the other side of the kitchen.

MS. JOHNSON: Yes, so go past the kitchen, go --

MS. WILBON: Oh, down this hallway?

MS. JOHNSON: Yes, so past the receptionist --

(Simultaneous speaking.)

MS. WILBON: Okay. Thank you.

CO-CHAIR CELLA: Just a quick question, do the SMP members on the phone have a hand raise function?

MS. JOHNSON: They do.

CO-CHAIR CELLA: All right.

(Simultaneous speaking.)

CO-CHAIR CELLA: Okay, all right.

MS. JOHNSON: She has taken care of that. Okay. Oh, yes, Patrick?

MEMBER ROMANO: One other question.

MS. JOHNSON: Uh-huh?

MEMBER ROMANO: At some point it might be helpful since we're going to be discussing materials that were distributed to other subgroups I'm not sure if there is a structure to the Sharepoint site, help find all of those materials?

MS. JOHNSON: Oh, yes.

MEMBER ROMANO: That might be helpful to give a bit of orientation to where to find the materials for other sectors.

MS. JOHNSON: Okay. Yes, very bluntly they are all under the Fall 2019 little tab, a plus mark that you see, and I don't even think they are in alphabetical order --

(Simultaneous speaking.)

MS. WILBON: -- numerical order by measure number.

MS. JOHNSON: It's numerical?

MS. WILBON: Mm-hmm.

MS. JOHNSON: I usually just do the "find" on the page and it will take me to that.

MS. WILBON: Yes.

MS. OGUNGBEMI: We can show you guys real quick.

MS. JOHNSON: Yes, we can bring that up.

(Off microphone comment.)

MS. JOHNSON: Yes, it's way down at the bottom. So all the measure documents are all at the bottom, yes, and then we are looking at Project, Fall 2019, and remember Sharepoint it has those little plus signs.

You have to remember that those plus signs will explode things for you. There you are. And then there are the measures. And Ashlie tells me they are in numeric order, they

are.

MS. WILBON: They are in numeric order.

MS. JOHNSON: Okay, all right. And like I say you can always do a "find" and just type in, usually I type in the measure number and it will take me to it.

So they are not split out into subgroups and things like that on the Sharepoint page.

(Off microphone comments.)

MS. JOHNSON: I was going to say I don't think they are in chronological order. Sharepoint is --

(Off microphone comments.)

MS. JOHNSON: No.

(Off microphone comments.)

MS. JOHNSON: No.

(Off microphone comments.)

MS. JOHNSON: Yes, I -- It surprised me when she said she thought they were because I didn't think they were.

MS. WILBON: Not in the order of the agenda, but they are in numerical order.

MS. JOHNSON: Are they actually in numerical order?

MS. WILBON: They are in numerical order.

MS. JOHNSON: Oh, they are, okay.

MS. WILBON: Yes.

MS. JOHNSON: Okay. So they are not in the same order that we will be discussing them, but they are in numerical order.

MS. WILBON: Yes.

MS. JOHNSON: All right, color me surprised.

MEMBER SIMON: Karen, can you just drill down into one of the measures because I am not sure I can get into -- I have a listing of the measures, but I don't seem to be able to access the actual material.

MS. JOHNSON: All right, okay. You should click on the blue link, so on the title, and then click. Again those plus signs --

MEMBER SIMON: Oh, okay.

MS. JOHNSON: -- will fool you every time. And the MIF, in case you guys, we have this odd -- We don't have a completely standardized naming structure.

MIF at NQF stands for Measure Information Form. That's where your specifications are going to live, and then we have the testing attachment, so usually the testing attachment is named testing attachment in some way or another, but developers will often re-name and put their own naming structure on that and we don't change it back.

The other things that are in these folders might be if they were say an appendix that the developer provided for you that might be in there as well.

I think we have the PA forms that all of you guys filled out, those are in there.

MS. OGUNGBEMI: Just the combined one.

MS. JOHNSON: Just the combined one, okay. So there is a combined PA form. You

probably named it something fairly similar.

MS. OGUNGBEMI: Right here.

MS. JOHNSON: Yes, I can't read that from here.

MS. OGUNGBEMI: It's named SMP combined SA/PA form and then the measure number.

MS. JOHNSON: Okay. So just so you know what the combined PA form is, all of you in the subgroups who did a PA we used the word combine function and combined all the information from the individual PAs that came through.

All of your comments are in there. They are in there and it will be noted by Panelist 1, Panelist 2, Panelist 3. So people will know who was on the subgroup but they won't know who made individual comments.

There may be a couple other types of things in there. If it's an e-measure then there is probably some HTML code and maybe some value sets, things like that in there.

It's pretty much what we were provided in terms of materials for you to look at plus the

PA form. Oh, and the other really important thing, the developer responses, so if the developers took advantage of their opportunity to give you additional information that is also in there as well.

Okay. Just so you know what's going on in this cycle, so we had a total of 55 measures submitted this cycle. That is really low for us this cycle.

So we're not exactly sure why we had relatively few measures, but it probably worked out, especially for your newbies, you didn't have quite so many measures to have to deal with this cycle.

So you guys looked at 22 of those 55, of those ten were new, the remainder were maintenance measures being considered for continued endorsements.

Okay. In terms of the types of measures, ten health outcomes, three composites, six intermediate clinical outcomes, and three PRO-PMs.

And then our initial results, which is always interesting to look at, and it will be interesting to see where we land after today and after tomorrow, 13 passed both reliability and validity.

Three of them did not pass either reliability or validity. So remember that in order to be endorsed measures have to pass both reliability and validity.

So three of them there was a no pass and then six there was some consensus not reached on one or the other, reliability or validity. So today we will be looking at 15 of those 22 measures.

So the ones that we are automatically looking at are the CNRs, the ones where consensus was not reached, automatically looking at those. We said that we would look at the ones that did not pass if the developers provided additional information and the developers for all three of the measures that there was a not pass on there they did provide additional information, so those

are on there as well.

And then, finally, measures where one of our panelists and/or staff pulled for discussion that also made our agenda. So, again, 15 of the 22 we'll be talking about today and tomorrow morning. Sherrie?

MEMBER KAPLAN: Can you define for all of us the definition of consensus not reached just one more time?

MS. JOHNSON: Yes.

MEMBER KAPLAN: Because some of it looked like it was a little bit difficult to understand.

MS. JOHNSON: It is a little difficult to understand. So you have for both reliability and validity you have the options of voting high, moderate, low, or insufficient.

So a high and a moderate rating together are passed, so in order to pass that criterion we have to have more, and I am looking at Ashlie and you tend to keep me honest, we have to have more than 60 percent with a high or

moderate gives us a pass.

MS. WILBON: I want to make sure that I have that right, too, because I get --

MS. JOHNSON: Okay. So more than --

(Simultaneous speaking.)

MS. WILBON: There it is.

MS. JOHNSON: Yes, that's a quorum. Yes, there, oh, great. You got it on here, all right. If it's between 40 and 60 percent that is CNR and that is inclusive, all right.

So if luck of the draw has it at 60 percent then it's CNR, it's not a pass, all right. The 60.111 would be pass. The less than is less than the 40 percent of the, in that bottom of either a high or a moderate.

We also have quorum. That means that we have to have enough people putting in a vote and that changes depending on the number of people in the subgroups who actually did what we would call the deep dive on the evaluation, okay.

MEMBER KAPLAN: One more clarifying question.

MS. JOHNSON: Uh-huh.

MEMBER KAPLAN: So in a six person issue three is half --

MS. JOHNSON: Yes.

MEMBER KAPLAN: -- but four puts -- So one person can change the definition of --

MS. JOHNSON: One person, yes.

MEMBER KAPLAN: Okay.

MS. JOHNSON: Yes. And it's a little tricky and that's actually -- We'll talk about that a little bit when we talk about how we're voting and stuff, right.

It's a little tricky, but that's where we are right now in our process. Patrick?

MEMBER ROMANO: Sorry to ask newbie questions, but --

MS. JOHNSON: That's fine.

MEMBER ROMANO: -- so the 33 measures that were submitted that didn't come here is that because they were process measures or for other reasons weren't considered complex measures?

MS. JOHNSON: Yes.

MEMBER ROMANO: And then linked to that, do we ever re-assess how the developers have classified their measure? For example, one of the measures that I think was submitted here that passed is a transfusion ratio measure that I would have considered to be a process measure but it was classified as an outcome measure by the developer.

MS. JOHNSON: Yes. That one is, the renal ones especially, what we do is we look at them as staff and sometimes we will actually have the developers change, sometimes we have a long discussion back and forth about whether we think they should change or not.

Those particular measures I think when they first came through I think we decided that they would okay as intermediate clinical outcomes, that's kind of where we thought they landed, and maybe they did land on an outcome outcome.

Some of things are really tricky and sometimes it doesn't matter in terms of our

criteria. If it doesn't matter then we will let the developer kind of pick which ones they prefer.

So there is some that are very obviously outcomes and very obviously process. Some of them are kind of straddling that line and we try to compromise and figure out the best way forward for those.

Okay. Can we go back to slide -- Any other questions? Then let's go to slide -- Oh, yes, John?

MEMBER BOTT: Yes, it was noted earlier that developers were given an opportunity to provide additional information and I see in the discussion guide at some points they did, but for some measures they did not.

Does that mean -- Well, the question is do they have the opportunity to come here today? If they did not present additional information up until now did they have the opportunity to come here today for the first time and present additional information?

MS. JOHNSON: We have said that the answer to that one is no because we wanted to -- I mean they could provide something a little different, maybe if you had a new question today that, you know, you didn't have before they could certainly do that if they are in the room or on the phone. They could do that.

Pretty much the only folks who do not provide extra stuff were a few that passed on both reliability and validity, so they didn't see the need to add additional stuff.

And, as a matter of fact, if those measures passed and were not pulled by us or by you guys then they are not the agenda so we're not even going to talk about them in these two days.

So we want to be able to allow them to give us more information, but we also want to have you guys to have at least a little time to think about and look at that information, so, no.

MEMBER BOTT: Okay. Great, thanks.

MS. JOHNSON: Any other questions?

(No audible response.)

MS. JOHNSON: Okay. So what's our process for today? It is a little bit different than what we have done on the phone, but not completely differently.

First of all, we will have staff introduce the measure. So we'll try not to be very long-winded with this, describe what the measure is and then we'll hand it over to the lead discussant to summarize key concerns, okay.

So there may be a lot of concerns on measures that there is usually a few that are the key ones, so that's the one that we'd like you to focus on and then we will ask other members of the subgroup to comment if you have other things that you want to add, so if you feel like there was something that you thought was really a key that your fellow subgroup member did not address.

Then we want to actually let the developers have two to three minutes to give you kind of synopsis of their response, so an initial response.

Then we want to open up the discussion to everybody on the full panel. So this is even if you were not a subgroup member for a particular measure you can still ask questions, you can still give your opinion about a methodology or whatever, whatever you want to do, the discussion is open.

A couple of notes here, a few people are recused. So that should be on your, we have an annotated agenda that should be listing all the people who are recused.

That means that not only can you not vote on a measure we also ask you not to discuss that measure. So you pretty much have to just kind of be a spectator on those if you are recused for a particular measure.

Now the developers, again, some of them may be in the room, some on the phone, they can certainly respond to any questions for panelists in this discussion.

So you would just say, you know, and, you know, developer, answer this question, and

let them know what you are curious about. And then we will have voting. Now this is -- I hope it's not completely confusing, we're going to try this.

We are going to ask the subgroup members, and these are the people who did the deep dive on these measures, to cast the final vote, okay.

Now that will be the vote in this room and, Yetunde, just make sure we are not going to not show the final count until everything is done, okay.

Those of you who did not, who were not part of the subgroup we're also going to ask you to vote, okay, but it's a shadow vote, all right. And I think I have -- I'll get to that in a minute. I have a slide that talks about that in a little more detail of what I am looking for, okay.

So we have some meeting materials for you. You should see on your, placed in front of you the annotated agenda and that tells you, we

don't necessarily expect you guys to remember what subgroup you were and all that kind of stuff, right, so it tells you which groups, who was in which subgroups, who the lead discussants are, and those of you who are recused, all right.

So that is all on the annotated agenda. It is printed out for you and we emailed that you I think on Thursday, I think. We also have the discussion guide.

So the discussion guide includes what we think is the pertinent information from the submissions. It's certainly not everything, right, because it's a two or three page summary, but the goal is to hopefully minimize the need for back and forth with the actual materials that were submitted.

Now that said, we can pull up the measure submission information for any measure, so if we need to do that we will. The measures are included in the same order as the agenda, so on the discussion guide everything is in the same order that's the way we will be going through

today.

We split it up first by subgroup, so Subgroup 1 we'll have your measures first, and then we'll go to Subgroup 2, and then within the Subgroup 1 measures we did it by rating.

So first we are going to discuss the ones that were CNR, or consensus not reached, and then we'll discuss the ones that did not pass and then we'll discuss the ones that were passed but not pulled, passed but pulled, and then I have passed but not pulled.

We're not going to discuss those, but those are actually on your discussion guide anyway. We wanted you to have the full information, okay.

Now Appendix B in the discussion guide actually includes all the additional information that was provided by the developers. So it's there in the appendix and it should also be linked to the appendix.

If you are looking at it electronically you should be able to just click

and go to the like section in appendix to look at the additional information that was provided.

MS. WILBON: There is also a bookmark in the PDF. If you open the bookmark tab you'll be able to see the outline so you can kind of click around and see that as well.

MS. JOHNSON: Okay, so the voting process. So we talked about this I think in one of our discussions as a group, and it might have been, it think it was probably before the new folks came aboard.

We talked about who should be voting. We're going to have everybody in the room. Do we allow everybody to vote even if everybody didn't do the deep dives or should we keep the votes just to the people who did the deep dives?

And there is pros and cons of doing each, right. So it would be great and it would be I think a stronger vote if we could allow everybody to hear, view, think, discuss, and then vote, and then we would have and end of 28 or something along those lines voting.

But we also know that people may feel that that's really unfair and that if you didn't spend quite a bit of time looking at the details of the measure it might not be fair for you to be able to cast a vote, okay.

So we tried to split it down the middle, okay. So we will have the official votes be like they were in the last two cycles, they are the subgroup votes.

So if there were six of you in a subgroup then our denominator for the voting will be six, okay. And it's also only the members of the subgroup who provided a PA who are eligible to vote.

So in a couple subgroups there were a couple of people who either couldn't do all of them measures or something like that, so the ones that you did the deep dive on and submitted the PA for you are eligible to vote in your subgroup.

And, again, those are the results that will be the official vote of the SMP. Now that said, we also want to know how -- We might not be

able to figure out how fair it is, but we wanted to get some information about a shadow, how others would have voted even if you didn't do that deep dive, okay, that's what I am looking for.

So we are asking you to vote via SurveyMonkey and we sent out --

MS. WILBON: I will send it right now.

MS. JOHNSON: Okay. She is sending out that link right now. So your email should get you that link to the SurveyMonkey and I'll tell you what the questions are on there in just a minute.

So if you did not provide a preliminary analysis for a measure then we would ask you to vote via SurveyMonkey, okay, and don't worry about it, you can look at the annotated agenda and you should be able to see which one you should or shouldn't be voting for, okay, or on.

Now the results from that shadow vote is going to help us determine if we should change

our voting structure in the future, okay. So some of the questions that we have on here, I'll just read them out, we want to know your name.

You have to pick which measure we are on, okay, you pick that out, and then we ask you this question, were you able to look at any of the materials for this measure.

So your choices are I did a thorough review of all measure materials, the discussion guide, the forms, the specs, all that kind of stuff, and anything additional, I did a thorough review of it.

And, again, we didn't expect you necessarily to do that, that would have been only if you wanted to do that beyond your subgroup. Your next option is I reviewed the information in the discussion guide only.

Your third option is I did a cursory review of some or all of the materials and then your fourth option is, no, I did not review this measure at all, okay.

And then once you tell us that we ask

you what your vote for reliability would be and your options are high, moderate, low, or insufficient, or N/A, this criteria was not reconsidered during the meeting, or I am not comfortable submitting a vote at this time.

Okay, so what we are trying to understand from you is how much extra work you guys did on these measures and given that amount of extra work how comfortable do you feel that you could make a fair rating after you have seen or read or heard whatever you seen or read or heard. That's the idea, okay.

So we'll see how this works. I have no idea, it will be really interesting to find out if overall the votes kind of track with what the subgroups end up doing or if they are very different, okay.

And Sherrie and Ron are looking at me and kind of smiling. I don't know if that means you don't want to know or it will be really interesting to find out or --

MEMBER KAPLAN: No, I was just

thinking that it would be better to have everybody vote and then understand what color their vote is, like -- something like that.

Because if you don't make people vote the probabilities that they will vote are probably limited and you're going to get sample sizes all over the place.

MS. JOHNSON: Yes.

MEMBER KAPLAN: Because if it was me I would consider --

MS. JOHNSON: You would have all --

(Simultaneous speaking.)

MEMBER KAPLAN: -- do at the NIH, you know, and then see what happens with that vote and see how much it would have moved somebody around depending on their level of comfort votes.

MS. JOHNSON: We did think about that a little bit. We actually have a comment box for you to add to in your SurveyMonkey so you can tell us maybe a little bit more you like doing that and -- Oh, I'm sorry, Terri, yes?

MEMBER WARHOLAK: I have a new person

question.

MS. JOHNSON: Mm-hmm.

MEMBER WARHOLAK: So there are two different voting platforms that I need to log into?

MS. JOHNSON: Yes.

MEMBER WARHOLAK: There is poll everywhere and that's if I am on the panel and then there is the SurveyMonkey if I am not, is that correct?

MS. JOHNSON: Yes. So you will be using both of them in the next day and a half.

MEMBER WARHOLAK: Okay.

MS. JOHNSON: Yes, and you'll be using mostly SurveyMonkey, right, as most of the measures you didn't do a deep dive on, unless you did a lot more homework than we assigned you.

MEMBER WARHOLAK: Okay. Thank you

MS. JOHNSON: Yes. Yes, and, Sherrie, we'll see how it works. It may turn out that we change it up next cycle, that we knew that -- Yes?

MEMBER WALTERS: I was just thinking that if it's me I'm lazy and I wouldn't like -- and it's easier for them to vote, I would do that rather than commit myself to making a vote and then trying to explain my --

MS. JOHNSON: Yes. Yes. We'll see how it works this time and we may actually do something very different next time. Yes?

MEMBER BOTT: I was just commenting you might want to study the reliability and validity of that process --

(Laughter.)

MEMBER BOTT: -- because I have my guesses as to what those are.

MS. JOHNSON: Yes. Yes. Yes. No, it will be interesting, we actually didn't know -- I will tell you what my gut feeling is, my gut feeling is that when hearing and looking at the discussion guide and hearing the discussion I feel like probably most of you will feel like you have enough information that it will be okay and fair and not unfair to cast that vote, but we

didn't want to make that decision for you.

We wanted to see if that was really true and then given that we wanted to see if there is any like systematic variation between those who did do the deep dive and those who didn't do it.

MEMBER BOTT: Yes. Apologies if somebody just said this, but I think you'd get more of an honest response for those of us who don't have to vote, or we weren't reviewing that measure if we did not have to type in our name, because I'm embarrassed of how few measures I reviewed --

(Laughter.)

MEMBER BOTT: But why ask for our name, what's it matter? You know, if 20-x of us did not review the measure who cares what person --

(Simultaneous speaking.)

FEMALE VOICE: That's true.

MS. JOHNSON: It's just -- We wouldn't have done anything -- Is that a have to --

(Simultaneous speaking.)

MS. WILBON: You don't have to -- Actually none of the fields are required. I mean we would like you to --

MS. JOHNSON: Yes.

(Simultaneous speaking.)

MS. WILBON: -- skip the name.

MEMBER BOTT: Santa Claus.

MS. JOHNSON: Skip it and do what Mitt Romney did and --

(Laughter.)

MEMBER BOTT: I just wanted you to get the best information --

(Simultaneous speaking.)

MS. JOHNSON: Yes. We don't need to know your name. We don't -- Yes, that's just kind of a typical thing we put in our surveys, so, yes we don't need to know your name. Just don't even do that, unless you don't care and you want to tell us that's fine, too.

(Off microphone comment.)

MS. JOHNSON: Is there anybody online,

on the phone who wanted to say something?

(No audible response.)

CO-CHAIR NERENZ: Maybe just one other, kind of in response to John, I haven't been on an NIH study section for awhile, but it was certainly understood at that time that not everybody on the study section read every grant, every page, every paper in full detail but we still voted and there was no shame in voting without having knowledge of the materials.

I might extend that to here, but we've been given a discussion guide, but also behind the discussion guide all kinds of materials. I would just declare, at least coming from me, I attach no shame to anybody who did not read all of the full detail materials for all of the measures in front of us.

That's part of the function of having a subgroup. So I am thinking that when it comes time for us to vote about things for which we are not part of the subgroup if all we saw was the discussion guide or if we just skimmed it or

maybe just right now we're looking through some of the materials, don't worry about anybody thinking you haven't done your job because there are a lot of materials here and that's why we have the subgroups.

MS. JOHNSON: I would want to be very careful of not to make you feel like we expected you to do the deep dives on all of them, because we do not, okay. We do not. If you want to that's great, but you don't have to and we don't expect you to.

And somebody on the line if you wouldn't mind muting your phone, we can hear some background noise. Thank you.

Any questions about that? I think it might be a little bit confusing at first, but I really think we were just a little nervous going straight to having everybody vote and using that one without having this kind of interim process to see how we think it's going to work.

MS. WILBON: And before we vote each time we'll queue you, we'll queue the group of

people who will be voting on which tool so you'll know, yes.

And even if you vote in the wrong place we'll figure it out and we'll work through it, so no worries. We'll guide you through the process.

MS. JOHNSON: We're also trying to be really careful and usually what we do when we do voting is you can kind of see the numbers kind of, as votes come in you can kind of see how the numbers are stacking up.

We have asked you today to hide that so that the people in the room are not going to influence the people who are doing the shadow voting, at least we hope that that's the case, right. So we think we can make that happen. Sherrie?

MEMBER KAPLAN: I just -- I won't interrupt you again --

MS. JOHNSON: That's okay.

MEMBER KAPLAN: -- but I just have one more question about the voting, the basis for

voting, because when we had these discussions before we sometimes raise issues that NQF has not yet presented to measures developers in terms of advancing measurement science and so on.

And so I want to make sure that in fairness to all the measurement developers that we are basing it on the criteria they responded to from NQF --

MS. JOHNSON: Yes.

MEMBER KAPLAN: -- and not what we are kind of starting to raise about maybe elevating the --

(Simultaneous speaking.)

MS. JOHNSON: Yes. So we will try. There is a couple things, one is I have a few slides just to remind people of some of the basic things, and then as we go through we will endeavor to make sure that we make it very clear.

So, for example, we've had discussions and I think many of you would love to see score level reliability testing across the board for all measures, right.

That is not a requirement at this point so we can't ding people for not having it unless their measure type says they have to have it.

And I'll get to another slide, but another one is showing a distribution of signal to noise reliability estimates, that is guidance, so it's guidance that you guys provided very, very recently.

It's too early to have expected developers to be able to bring that and so we can't hold them to that. So we will be very, we will be paying attention, all of us, on the inference and trying to make sure that we don't let any of that go through.

If you have a question please answer because we do want to be very fair. We don't want to hold people to something that is not one of our requirements. And, Jack?

MEMBER NEEDLEMAN: Yes. I had a slightly different perspective than Sherrie on this.

I agree with you, you can't hold people for documentation that was not required, but the standards have been evolving and the standards that might have led to approval in the past are not necessarily the standards that will lead to approval right now.

And absence of convincing documentation that something has been met has led, at least some of us, to describe measures as less than reliable or less than valid and that discussion I suspect is going to continue during the day.

MS. JOHNSON: I agree. And I will also say that, you know, you guys know things to look for and it may be in the past others didn't know to look for, all right.

So just by definition of having this panel we are going to be looking a little more rigorously. All right, there we go. Just a reminder about our ratings, if you want to give a measure a high rating remember that score level testing is required, okay.

Otherwise, and it would be a moderate rating would be the highest, a measure may be eligible for a high but the sampling method, the results, something like that may make you choose a moderate.

So we do have those algorithms that the algorithms tell you the highest possible rating. It doesn't mean that if it falls its way through on a certain place in the algorithm then it has to have that rating, okay.

So for a moderate it's the highest rating, it's the highest rating eligible. If only data element is tested, is provided, or if face validity testing is conducted.

So if all you have in front of you is face validity then you should never be giving a high rating, okay. And the same thing, it could be eligible for moderate but because the sampling or the methodology or the actual results you may choose low instead. You don't have to choose moderate.

Low is used primarily if the testing

results are not satisfactory. So if we assumed that the methodology is correct and that the data and the testing sample is reasonable but the numbers just don't satisfy you so that would be low.

And that's different for insufficient. So you would use insufficient when you really just don't feel like you have enough information to assign one or the other leads, just something is just not clear enough that you feel comfortable assigning a lead.

Okay. Just a reminder, none of this is new to you guys. Testing requirements do vary by type of measure. So for health outcomes, intermediate clinical outcomes, cost and resource use structure process all of them are, we have different requirements.

For both reliability and validity NQF requires either data element testing or score level testing, okay. So that's, there is final outcomes, intermediate outcomes, cost resource use structure process, so it could be either or.

We would prefer both but we don't require both, okay. So you can't take them down if they don't have both. It impacts the ratings, as we have already described, and the exception is the face validity for new measures we do accept it.

So on your discussion guide we tell you right under the name of the measure we tell you if it's a new measure or a maintenance measure.

If it's a new measure and they are providing face validity that's okay, okay, unless it is one of these other kinds of measures. So this is where it gets a little tricky.

The next one is probably something that you guys don't like very much and maybe something that we might want to change at some point.

If data element validity testing is provided we do not at this point require additional reliability testing, okay. So then what do you do?

Well you figure out what you would have did as the rating for validity, the validity testing, you look at that and apply that to reliability.

Now we still do see this, it happens on occasion. It's not nearly as common as it used to be, okay, so it's self-correcting a little bit, but that's just something that we've had for a long time.

Again, it might be something that in the future we will change, but right now that's one of our kind of rules.

CO-CHAIR NERENZ: What does the word additional mean? Does that mean, for example, you don't need measure score of reliability?

MS. JOHNSON: Right.

CO-CHAIR NERENZ: Does it mean you don't need any reliability?

MS. JOHNSON: You don't need any. So if you have shown the validity of the data elements then we don't make you do anything else as long as it's one of these kinds of measures.

Larry?

MEMBER GLANCE: Just a quick comment. So I think all of us understand the structure that you are presenting. I think that some of this probably a lot of us are uncomfortable with saying that a measure is scientifically acceptable if the only thing people have looked at is data quality, meaning data reliability and data validity.

Luckily, I think most measure developers provide us with information on score level validity and score level reliability, but I think in the rare instances where the only thing we hear about is data quality, you know, even though it's sort of like we're changing things on the fly a little bit I think it would be very difficult for me to say, yes, you know, you've told us about data reliability, you've told us about data validity, this is the only thing that NQF asked you for and we're okay saying that this is scientifically acceptable.

And I also know that having spent

several years on the standing committee for re-admission measures I think people sort of understood this structure, but at the end of the day if it was a risk-adjusted outcome measure and the risk adjustment model was no good, so there was a problem with score level validity, we just voted it down regardless of what this algorithm said.

I think it's important to think about that when we are looking at the measures today because honestly I think in this day and age if the only thing we have on a measure is data quality information I don't think we can really pass it as being scientifically acceptable regardless of what this algorithm says.

That's my opinion and I think practically speaking I think that may be what a lot of us employed when we were looking at these measures.

MS. JOHNSON: I think what -- So a couple things on that. First of all, we would like you to apply the criteria and just vote

whether you think it meets our requirements.

So I know that it's going to make some of you uncomfortable in certain situations, but don't hold them higher than our requirements, okay.

Now that said, most of the time, almost every time, the measures that come to you guys, the conflicts measures, almost always have these other things so you're not going to get into this kind of uncomfortable situation really often.

That's not necessarily true for the process measures and/or structure measures that you guys aren't seeing, okay. Now we can and I expect that we will be putting forward in March to our CSAC, who oversees our criteria, we will probably be putting forward some of your suggestions, such as this, should we raise the bar of what we're asking and at that time we will kind of talk about the pros and cons, the ins and outs, and see what the CSAC allows us to do.

So, yes, I think you might be in a bit

of an uncomfortable situation on a few of your measures and, again, we really need consistency to the extent that we can get it, so please try to use our criteria as the basis of your votes.

CO-CHAIR NERENZ: And just again as we get into this now, because I know it's about time, as each measure comes up why don't you just remind us, staff or somebody, what are our requirements for this measure and what must it have and then we can just keep that in mind as we go through.

MS. JOHNSON: Yes, we can do that. All right, there we go. Composite measures, we have a couple different kinds of composites, kind of the traditional ones where you have data aggregated at the measure entity and then -- Sorry, I didn't say that right. Data that are, the components are aggregated at the measure entity and then aggregated again as a composite.

We also have all or none measures. So those are what we mean by composite measures and those actually have an additional criterion that

we'd have to look at and that also is must pass.

But that does not, we do not count the measures that are based on multi-item scales. So even though the word composite is used in that scenario that's not how we are using it at NQF, okay, so we will do what Dave has reminded us and remind you again of our composite definition.

Slow level validity testing is not required for composite measures until they become a maintenance measure in which case we do expect it, okay, and I already talked about the additional --

And instrument-based measures, which we don't have as many this cycle as we did last cycle. Last cycle we had tons of them. This time we don't, but this is one where we actually are expecting both levels of analysis and we've had this discussion many times what do we mean by data element, we just really want to know that the relevant items in the instrument are reliable and valid.

And I know, Sherrie, we haven't

probably quite got that perfectly yet, but that's the idea, but then we also want to go beyond just how good is the instrument, we want to know how good in terms of reliability and validity the actual performance measure is.

And then another reminder, and this sometimes can throw people off, we allow people to put multiple performance measures under the same NQF number, okay.

So what that means is we might be discussing Measure 1234 but if it is an instrument-based measure there may be ten individual PRO-PMs or other kinds of PMs underneath that number.

All of those you have to think about those separately. You could end up, if you need to you could provide separate scores or separate ratings for those if you feel like we need to, and some of them could actually pass and others not pass, okay.

So it's more of a convenience that we have allowed people to put them under the same

number, but we do consider them separate measures.

All right, eCQMs, eMeasures, testing for a more and more EHR system is required. This is I think the only definite that we have in terms of testing.

Testing is reliability testing that is not required if based on data from structured data fields, but if based on unstructured data fields then we would require both reliability and validity testing.

And as of the summer of this year we actually now require data element validation for eCQMs, and that is new.

Okay, a few more reminders. We are almost done. Testing must align with specifications. So this is not something new but it is something that we have been more rigorous in upholding and we are particularly interested in level of analysis and minimum sample sizes.

So what we are saying here is if a measure developer says that this is specified for

two levels of analysis we expect separate testing for both of those levels of analysis, okay.

And similar for the minimum sample sizes, if there is no minimum like threshold each provider has to have at least say 25 patients or something like that, if that's not specified then we expect the testing not to have limited to just 25 or more, something along those lines. We want those to align.

I have already talked about the performance measures and how there could be more than one under one NQF number. It's very obvious when you are talking about PRO-PMs all based on one instrument, but it can happen with other kinds of measures as well.

Okay. For risk-adjusted measures just remember that inclusion or not of certain risk factors cannot be a reason for rejecting a measure, so you should certainly talk about any concerns that you have, but, you know, whether or not one particular factor or set of factors is or isn't included can't beat out why you would take

a measure completely down.

That said, the information about discrimination calibration or overall methodology of risk adjustment can be grounds for taking a measure down or not passing a measure.

For all measures we want to make sure that we understand the specifications. And, finally, empirical testing is required at the, or is expected at the time of maintenance evaluation.

And, again, I've already mentioned this, this is the face validity thing. If face validity is still being used then there should be a justification for that.

And, Patrick, do you have a question?

MEMBER ROMANO: Yes. On the last slide, could you just go back? I just wanted to clarify this last bullet.

So, of course, there are many examples of this last point that I just wanted to make sure I have the right understanding, which is that NQF is focusing on measures that are used or

usable for both quality improvement and accountability applications and, therefore, it's often the case that, for example, a composite might have a set of component measures and those components might be used internally.

They might be reported as part of a dashboard from a registry sponsor for example, but if there is one measure that is the composite that is actually going forward as the accountability measure that's what we are voting on, is that right?

MS. JOHNSON: Yes.

(Simultaneous speaking.)

MEMBER ROMANO: So we are not passing each of the component measures separately?

MS. JOHNSON: No. Not in the composite measures, no. Now occasionally --

MEMBER ROMANO: But even in cases where there might be -- I think we had a case here where there is a bunch of measures that are being proposed essentially for internal purposes as different ways of looking at the same problem

but there is one measure that is the accountability measure. So shouldn't we focus on that measure?

MS. JOHNSON: Yes. So the developers will be telling you occasionally there will be a composite measure or something along those lines and they may say we're only putting forward the composite for a potential endorsement, so that's what you should be paying attention to with our criteria.

Now sometimes when you are evaluating composite you do have to think about the individual components and pieces of those, right, but maybe you are not, especially for reliability you might not be as interested in the components.

Sometimes even with composites they will say I'm actually interested in having individual things looked at for endorsement and the composite, so they should be telling you and they should be very clear are we looking at three performance measures, are we looking at one, are we looking at ten, are we looking at one.

MEMBER ROMANO: And would staff then force them to split those into separate number or not necessarily?

MS. JOHNSON: Not necessarily.

CO-CHAIR NERENZ: Actually one thing that might help clarify, you know, a composite is one thing. You basically take a set of measures and you mathematically combine them and you get a composite measure and then that is up for endorsement.

The CAHPS survey is an example in any of the setting it's used. It is one survey and it gives one number at NQF but from it are derived several measures.

They are not composites in the definition we are using here. They are separate measures. I don't think we have any in front of us today, but in the prior cycle those of us in the subgroup who might have a hospital CAHPS survey that generates six measures and we have to evaluate all six of them but they have one number, but that's not a composite.

CO-CHAIR CELLA: Matt and then Sherrie.

MEMBER AUSTIN: Yes, I mean it's sort on this note of having multiple measures under one number. There has been a couple of measures through the rounds where it feels like I am being asked to evaluate different measures under the same form.

Can I just maybe get some clarification on why NQF chooses not to split those out into separate measures? It just feels slightly confusing trying to assess sort of multiple measures that are under one number.

MS. JOHNSON: Yes. Yes, and in that case I think it is a little bit unfair that the onus is a little bit on you that if you feel like certain pieces need a certain rating you have to tell us that, you know, because we don't give you that, kind of we don't split out and give you a special form.

It really comes down to us trying to be, I don't know what the right word is, us

trying to kind of meet the developers in the middle who don't want to fill out ten forms for each of their, for example, hospital CAHPS measures, actually 11, because so many of the things, not just -- Reliability and validity, the methodology and the sampling, the testing dataset would usually be the same.

The numbers, the results will be different, but other things like evidence, the usability, the feasibility, all those kinds of things are the same and they really don't like copying and pasting so we have allowed them to put things on one form.

So it's not optimal. It makes our counting off, too, right, because we'll count and we'll say, you know, we have one hospital CAHPS, but we really don't, we have 11 measures that were based on the hospital CAHPS.

CO-CHAIR CELLA: Sherrie?

MEMBER KAPLAN: I just wanted to clarify because this terminology confusion it still confuses me. An instrument the way it's

being considered is a data collection method, right?

MS. JOHNSON: Yes.

MEMBER KAPLAN: So that it's a survey, for example, in the case of HCAHPS. HCAHPS is if you are measuring the whole construct of overall quality of care it is the whole score.

But if you are measuring doctor/patient communication it's meant to be used at the provider level out of that then that is a measure of a construct, doctor/patient communication, in that case it has three items and the precision gets really at issue then.

So, you know, if for example it wasn't HCAHPS and it's only the sort of overall concepts of being measured with that data collection instrument, like lumbar pain and other kinds of things that are related, but the whole instrument has multiple different measures on it that aren't necessarily part of an overall construct then it gets confusing.

You know, you would not in that case

evaluate the instrument. You would not evaluate the whole thing because it's got multiple different and independent parts.

MS. JOHNSON: Yes. You just --

MEMBER KAPLAN: So NQF will never ask us to evaluate an instrument with things that are unrelated measures that are unrelated on it.

MS. JOHNSON: Right. We would only ask you to be thinking about the things that go into that particular measure.

So they might have, and as a matter of fact we have an example of that today where it's a measure that has several different domains in there but they are only putting forward one of those domains for endorsement.

So we're not asking for all the information about all that other stuff. We're only interested in that one.

CO-CHAIR NERENZ: Okay, we got to keep going.

MS. JOHNSON: All right. I think that might have been it.

(Multiple nos.)

(Simultaneous speaking.)

MEMBER NEEDLEMAN: If we move to the next slides as Patrick brought us back, I want to talk about, I need some clarification on the risk adjustment statement because we've had a lot of discussion in this committee that the risk adjusters, whether they are doing what we think they are doing and whether they are doing it appropriately.

So the statement about inclusion or not inclusion of certain things seems to run counter to the discussion we've had about socioeconomics, say certainly there is a lot of presentation in there.

When we have looked at it as individuals collectively we've been wondering whether the variance explained by the risk adjuster is about the right level it would expect given the variance in the thing being studied and the precision of the risk adjuster.

We have asked how much the rankings or

the scores change as a result of the inclusion of risk adjustment. All those I assume fall into the calibration issue.

But there is also an issue of over adjusting and that implicitly, if not explicitly, deals with the issue of there is some things in this risk adjuster that shouldn't be in there and that seems to directly conflict with the guidance you are giving us on don't comment on individual elements of the risk adjustment, so I would like some clarification.

MS. JOHNSON: Yes. We want you to be very comfortable commenting on any of that stuff, so please do if you have any concerns about any of that definitely comment.

We are just saying we don't want you to rate as low because of that only, okay. So don't take it down because of that, and the reason is that we actually on these things, some of these things are very much clinically, there is content and clinical judgement that is needed for that and while many of you have that we don't

necessarily, we're not asking you necessarily to put on your clinical hat for that.

We actually want that decision to be a standing committee decision.

MEMBER NEEDLEMAN: And since I don't have a clinical hat, although as an economist I practice many professions without a license.

(Laughter.)

MEMBER NEEDLEMAN: I have in my comments said the standing committees take more of a look at this.

MS. JOHNSON: Yes.

MEMBER NEEDLEMAN: That's the sort of guidance you want from us?

MS. JOHNSON: That's what we want, yes. So we don't want you to not talk about it, we just don't want you to fail. We want them to have the opportunity to hear your concerns and then take it from there.

Okay. There is only one more slide and I think I have already gotten to that one.

CO-CHAIR NERENZ: All right.

MS. JOHNSON: And then, Zhenqiu, I will get you in just a second. We do have a little bit of additional guidelines in terms of asking for more detail about methodology reporting only and one overall statistic for signal to noise reliability, giving a whole lot more information about the construct validation methodology and what you are trying to do.

This is all new guidance, we've talked about this. Don't hold our developers in this cycle to these things now, this is still relatively new.

It's not that you don't want it, it's just that we don't want to take them down for these at this point, okay.

CO-CHAIR NERENZ: Zhenqiu?

MEMBER LIN: So to follow on Jack's question, so for example this measure about mortality, and then you view it through in hospital complication as research has been variable and then we wouldn't agree that's a right thing to do, right?

If that happened to me that should be grounds for a rejection.

MS. WILBON: I'm sorry, Zhenqiu, can you speak up just a little bit --

MS. JOHNSON: Yes, I --

MS. WILBON: -- I couldn't hear you.

MEMBER LIN: So if the measure is a mortality measure and then you include the hospital complication as risk adjustment variable which is a direct result of quality of care, all this is going to predicable, right.

But in that case to me like -- this you include that as risk variable that should be ground for rejection I would imagine.

MS. JOHNSON: You know, I see what you are saying and I think technically it should be. I think we would probably ask you not to take it down if that's the only thing, but be very clear that that is a huge concern and just say, you know, that's not part of risk adjustment, you're not supposed to do it that way, and that sort of thing.

CO-CHAIR CELLA: So I'll just add, I thought of examples like that also, Zhenqiu, and I kind of took comfort in the sub-bullet that says concerns with discrimination calibration and overall method of adjustment are still grounds for rejecting.

So just because -- The way I took it is just because you might think and maybe even feel strongly that a certain variable should be included you have to have a reason that relates to the overall adjustment as a result of it.

So I think there is a clause for making the case you just made. Does that make sense?

MEMBER LIN: I mean I would defer it to clinician in this room, right, you've got in quality of care and then complication happen during part of a hospital stay, right, do you think that should be included?

MS. JOHNSON: Why don't we hold off and if that --

CO-CHAIR CELLA: Let's see if it comes

up, yes.

MS. JOHNSON: -- specifically comes up then we'll just talk about it as part of the discussion of that measure.

CO-CHAIR CELLA: And as we try to -- And we're all sort of, hopefully Dave and Dave are getting aware that we are moving past our time and we always think there is plenty of time to go through the overview and then we get a late start.

So I just want to say, you know, how when you are in a group and you are new to a game and somebody explains the rules and you don't think you really understand the game and they say well just start playing, it will become clear as we go along.

I think we need to just start playing the game. The NQF staff is really good at reminding us if we're going off the guidelines. And, also, Ashlie just sent around, because it looks like most of you are on email, just sent around the slides that have a lot of these

principles in there.

So the slides are available to you to refer to during the discussions and let's start playing.

MS. JOHNSON: Let's start.

CO-CHAIR CELLA: All right.

MS. JOHNSON: All right. Yetunde, you might want to take us, this -- Okay, a voting test. Do you want to walk us through our test?

MS. OGUNGBEMI: Yes. Now I do have a question, do you want them to see the voting at all?

MS. JOHNSON: We need to see the final score once everybody is done. We have to --

MS. OGUNGBEMI: Okay. Where would you like it projected?

MS. JOHNSON: What's our choices?

MS. OGUNGBEMI: There or there.

MS. JOHNSON: There.

MS. OGUNGBEMI: Okay. Okay, so during the test vote you can -- Oh, first of all, did everyone receive the email with the voting

instructions?

MALE VOICE: Well there is two emails with voting instructions, which one?

CO-CHAIR NERENZ: Which one?

MALE VOICE: 7:50 a.m. or 9:07 a.m.?

MS. WILBON: The 7:50 one.

MS. OGUNGBEMI: Yes. So the link is wholeev.com/nationalqual661 and we'll go to that voting link. The voting will become active only when I press the activate button.

So it will show like a blue or gray screen currently if you aren't logged into that poll. All you have to do is type in your name and the survey will come up.

MS. JOHNSON: Okay, and everybody is doing this test, right. This is just to make sure you know how the in-room voting goes.

MS. OGUNGBEMI: So the voting is active now. During each vote I will read the question and the options. There will be four options for all of our votes today and they will rank high, moderate, low, or insufficient.

Of course for this test it's not required for that. And so this poll here will show responses, as soon as I click this it stops the responses from being shown.

So you can see them now and then once we say the vote will close I can show the responses in the room if that's what we would prefer to do.

MEMBER NEEDLEMAN: And does only our first vote count our only our last vote?

MS. OGUNGBEMI: The last vote.

(Pause.)

MS. JOHNSON: All right, everybody. okay with this measure?

PARTICIPANT: This is the subgroup, right?

MS. JOHNSON: This is what you'll be doing if you're part of the sub-groups, okay. If you're not part of the sub-groups, you'll be doing the SurveyMonkey that we're not going to do a test run.

All right, Andrew, I think you might

be up. So we are going to move to Measure 2456, Medication and Reconciliation, Number of Unintended Medication Discrepancies per Patient.

And Andrew is going to lead us through. Again, just a reminder, Andrew will do his thing. We'll hand it off to Jen, on the phone. And then we'll let the subgroup add any concerns they maybe Jen may bring up. And then we'll ask the developer to respond. And then we'll open it up to anybody.

MS. WILBON: Can we just announce to you that the subgroup, we're starting Subgroup 1 measures. That includes Dave Cella, Sherrie Kaplan, Zhenqiu, Jack Needleman, Jen Perloff, Christie Teigland, Terri Warholak, and Susan, oh, excuse me, Susan is not here. I'm sorry, and Terri Warholak. So just, you guys are on deck.

MR. LYZENGA: Okay. I was trying to get this mic to work. I'm a hopelessly quiet talker so I'll try to project.

So this is a measure of the quality of the medication reconciliation, an interesting

measure unlike many of the med rec measures we often see here. This is not accessing any particular process or documentation or attestation that some processes occurred but, instead, it is looking at the actual number of unintentional discrepancies on the medication list per patient.

The measure requires a trained pharmacist to do a review of the medication, a patient's medication list on admission and then at discharge. And then that is compared to what was actually taken during the patient's care and the number of, some additional analysis and reviews done to see whether those discrepancies were intentional or unintentional. The number of unintentional discrepancies per patient is then the basis of the measure.

So this is sort of like an intermediate outcome of sorts. It's not sort of a pure process measure. It's not exactly an outcome either. The measure, the developer recommends sampling 25 patients per month or

approximately one patient per weekday.

In terms of the testing, we had a data element assessment done. There was an inter-rater reliability assessment done where two study pharmacists independently collected the medication histories for 19 randomly selected patients and calculated the percentage of agreement.

That was one part of the reliability testing. The sort of raw rate of agreement, there was 77 percent. And that was looking at where there was complete agreement in medication dose, route, and frequency across the two assessments.

The developer also assessed the reliability of the discrepancies for its system, so whether, you know, there were, the independent reviewer got the same number of discrepancies as the original review.

In that case, they got 91 percent agreement and did calculate a kappa, both for admission discrepancies and discharge

discrepancies. The kappa there was 0.64, indicating in the interpretation of the developer's substantial agreement.

For validity there was also, well, to put that in straight validity, and in some sense this is sort of a base validity assessment, the developer did note that the literature shows that pharmacists made more accurate medication histories than either nurses or physicians and essentially suggests that the pre-admission and discharge medication history taken by trained pharmacists is itself, in some sense, the gold standard.

And they sort of claim that they don't have any real gold standard to which to compare this to and, you know, assert that that is, in some sense, a face validity in itself that this is the gold standard process.

They did, in addition, provide a bit, well, they also provide some materials to show how these expert pharmacists are trained and to show that the process is transparent and

systematic.

They also provided some data showing that, in their study sites, 9 of 17 study sites had significant improvement in their discrepancy rates in the last six months compared to the first six months, and that those that did improve their rates had a greater increase in the number of patients receiving recommended patient level interventions such as this face to face collection of the medication history.

So I think that's intended to be a sort of score level validity. It's not what we typically see but showing that there is some correlation between best practices and improvement in the rate of discrepancies that occurred here.

We did get a passing rating for reliability, consensus not reached on validity. Oh, I should add that this measure is not a composite or instrument-based, so we don't have any additional requirements. Data element validity, and reliability testing is acceptable

for this measure.

The face validity is something that we actually expect in careful validity testing at the time of maintenance. This is a maintenance measure. But we do accept face validity with a rationale from the developer. I think that rationale is, again, basically that they didn't really have something to compare it to. Because they are using the gold standard itself as a basis for the measure.

A few of the questions that came up as part of the subgroup reviews were whether there ought to be a minimum number of patients required to calculate this measure.

Some, you know, suggestion that the standing committee of content experts should review the training and instruction materials to make sure that those are indeed adequate, that it is, in fact, a gold standard medication history being collected.

Some concerns about the reliability testing, the result of 77 percent agreement is

not terribly high, necessarily. And then the testing of the discrepancies, some suggested that they should be based on more cases. I think there were only four cases that actually were the subject of that analysis.

In terms of validity ---

MS. MUNTHALI: Andrew?

MR. LYZENGA: Yes.

MS. MUNTHALI: I think we can let the lead discussant, maybe go through some of those items.

MR. LYZENGA: Okay, sure. Maybe I'll just stop there.

MS. MUNTHALI: Yes, that would ---

MR. LYZENGA: So I think that's Jen on the phone. Are you there?

CO-CHAIR CELLA: Dr. Perloff? On mute maybe? We'll continue on. She's not listening, I guess, right now. I thought she was signed in earlier.

MR. LYZENGA: All right, so in terms of validity, some of the concerns were that

excluding, there's an exclusion for patients who are unavailable to be seen by a pharmacist or who decline to talk to the pharmacist.

There was some concern that that did produce bias and, you know, to see if that could be, or it could be used as an excuse for not getting data on difficult patients, and wondered if, to address this, there should be a time limit set such as those discharged in six hours or less should be excluded from the ---

With regard to the validity testing, reviewers noted that while the measure does appear to have face validity from a common sense perspective, there was no systematic assessment done and potentially a lack of data provided to support the assertion that this measure can reliably and validly identify discrepancies, even if it does use what is considered as the gold standard for gathering medication histories.

So sort of the questions here are whether the developer's rationale for case validity of the measure is adequate, and his use

of what is considered the gold standard practice for collecting medication histories, is sufficient as a demonstration of validity.

And then, you know, considering this additional data, suggesting improvement in measure performance being associated with consistent implementation of best practices, this does provide additional support for the measure's validity.

Since this past unreliability, are we really just focusing on the validity discussion here?

CO-CHAIR CELLA: Yes.

MR. LYZENGA: I guess we can open it up for discussion at this point, again focusing, if we can, on validity.

MS. WILBON: We'll open it up to the subgroup at this time.

MR. LYZENGA: The subgroup.

MS. WILBON: Yes.

MEMBER KAPLAN: Okay. So I was a little confused about enumerating. Because if

the number of patients, and I need some help here, I can get it from the developers. The enumerator statement, the number for each sample patient in the denominator group, the total number of unintentional medication discrepancies, it struck me that the more medication you're on, the more opportunities you have for discrepancies.

So I wasn't clear from the adjustment strategy that that was somehow managed. Because what if some facilities have a whole bunch of CHF patients, for example?

MR. LYZENGA: I do believe they accounted for that. We can ask the developers if they're on the line. I think the, it's actually discrepancies per medication to account for that question.

MEMBER KAPLAN: But then the enumerator is not the number of medications.

MR. LYZENGA: I think the enumerator is the number of discrepancies.

CO-CHAIR CELLA: The denominator is

the number.

MR. LYZENGA: That denominator is the number of ---

MEMBER KAPLAN: It's the number of medications or the number of ---

MR. LYZENGA: Is the developer on the line? Maybe they could provide us some clarification on that.

MEASURE DEVELOPER: Yes, this is Dr. Schnipper. I'm the measure developer. How are you?

MR. LYZENGA: Good. We can hear you.

MEASURE DEVELOPER: Good, you can hear me okay?

MR. LYZENGA: Yes.

MEASURE DEVELOPER: Yes. So the measure is the number of discrepancies per medication. Because the more medications you're on, the more opportunities there are for error. And so we look at discrepancies in admission orders and discharge orders. So basically the the maximum number of discrepancies per med is

two. But the meds can be wrong on the admission order and can be wrong on the discharge order. And so that basically adjusts for the number of medications that the patient is on.

We see no other need to adjust for anything else in our studies. And now we've done two studies at teaching hospitals. But I'm happy to answer any more questions about that particular piece. Thank you.

MEMBER KAPLAN: So I have a follow-up question on that. Because if a facility has a small number of patients on a huge number of medications, and the precision of the estimates then is, so that's just a small number of patients.

So smaller facilities might have fewer opportunities to have discrepancies, because they have smaller numbers of patients? Help me understand what the relationship of facility size is to this.

MEASURE DEVELOPER: That's a great question. And I think the question came up a few

minutes ago. I did sort of a minimum number of cases to provide validity. You know, in our studies, we did 22 patients per month. And that was enough patients, I think, to get a fair representation of what's going on in your hospital.

We try currently doing 20 per quarter. I think it's questionable if that is enough. So I think that's, you know, something that we should d definitely talk about offline. But a hospital has to be big enough to provide enough of a sample of its patients that the measure is an adequate representation of, you know, of their quality of care.

If a hospital is too small to provide that number, whatever that number is that we agreed to, then it would be too small to be able to carry out this measure. So, you know, I would pick a hospital with, you know, 20 beds or something like that, it might not be.

And I don't know if you ever exempt hospitals of a certain size, certain measures,

but there would be a concern, under those circumstances, that you wouldn't have enough of a sample size.

MEMBER KAPLAN: Got it. So just two more quick questions. One is your follow-up data from the MARQUEE study, MARQUEE II study, shows a significant improvement in half of the samples -- of the site studies, but no change or an increase in medication discrepancies in the remaining sites. And that's what you're citing as validity.

My concern is that this measure's been around since 2012, it looks like. And so in the meantime, we've accrued a lot more opportunities for construct validity. So what would you assume that people who have a large number of medication discrepancies would have? For example, more likely to return to the hospital for readmission within 30 days or so on.

Have you considered, first of all, with respect to only half of the studies responded, or it looks like they're improving and

the rest either didn't change or they got worse, is that good evidence of validity?

And then the second question is why would you not then think about some other construct validity type variables to test this against?

MEASURE DEVELOPER: Right. And those are all good questions. So across the 18 hospitals, the reduction in discrepancies, on average, is very strong and consistent. It's almost a straight line down from about three discrepancies per patient to about one discrepancy per patient.

Again, there's constant improvement over the time of the study. Comparing the third time period to the first time period is an insensitive way to measure. But it was an easy one to do, you know, for this particular measure. But overall, the effects of the MARQUEE II study were very strong.

We're about to submit that work publication in JAMA this month. So there's that

piece. We are in the midst now, in terms of concept validity, if we're comparing improvements in discrepancy rates with reductions in hospital length of stay and in readmission rates.

And that's fine, it's just that those, you know, announcements aren't completed yet. But, you know, it's a large study with a lot of people. And, you know, the problem with re-admissions and length of stay is that neither were planned analyses but were added on.

And they don't always do risk adjustments the same way. Not all of them prescribed have been observed over expected lengths of stay. Some hospitals that had more selection of who got the intervention versus who didn't so they could be bounded by indication, given who got the interventions versus who didn't, when you want to compare those who did and didn't.

So it's more complicated than just that. So we are trying to do concept validity with both in-hospital length of stay and

readmission, we just don't have the data for you. I apologize.

CO-CHAIR CELLA: Larry and then Jack.

MEMBER GLANCE: So I just wanted to follow-up with a comment on what, Chair, what you just made. So if you have, there's no risk adjustment for this measure, correct?

PARTICIPANT: Right.

MEMBER GLANCE: It seems like there's a big ---

MEASURE DEVELOPER: Yes. We, okay, go ahead. Sorry.

MEMBER GLANCE: It seems like there would be a big difference where, if you have the patient who's on 25 different medications, it's much harder to achieve successful reconciliation for that patient versus a patient who has two different medications.

And the way the measure's currently structured, you could have the same outcome for 25 patients who are each on one medication versus one patient who is on 25 different medications.

So in terms of score level validity risk adjustment, I think there ought to be risk adjustment in this measure. I think it's very difficult on the face of it to say it's okay to put this measure out in the wild without any risk adjustment whatsoever.

MEASURE DEVELOPER: So a couple of comments here. Those are all completely valid comments. You know, I would say that we do discrepancies per med, per patient. So one patient with 25 is not the same as 25 patients on one medication. You would still be gathering data on 25 different patients. And each one of them would have the number of discrepancies per the number of meds that they are on. And then you average that across all of your patients, so just to make that clear.

In our studies, we do a risk adjustment for a few basic things, including age, and including the understanding of the patients, the patients' overall understanding of their medications as determined by (telephonic

interference) medication.

And certainly we have some criteria such as high, medium, and low. And the question was simply, you know, in terms of ease of operationalization, whether that was necessary or not.

But one point I would make is that, you know, so Leapfrog is using our measure. There are now thousands of hospitals that are doing it. Right now, as I said, at a low level.

The question is really is the hospital being compared to a different hospital, or is the hospital being compared to itself?

I think there's no doubt in my mind that comparing a discrepancy rate over time within a hospital is a valid measure whether you're improving care in that hospital.

I think the broader question of I've got an absolute discrepancy rate of X, and you've got a discrepancy rate of Y, does that mean you're a better hospital than I am? And I'm often not convinced that that may necessarily be

true.

I think the purpose of this measure has been to drive internal quality improvements to see improvements in your own hospital over time. And for that, I think it's clear that it does that and it really does compare. It does tell you, you know, if your hospital's improving your processes of care or not.

So, you know, I would be happy to talk offline and explore ways to do more risk adjustment to compare hospitals to hospitals, you know, one to another, and be sure that you really measured whether, you know, you're providing better care than another hospital is within a hospital. It definitely is measuring whether improvement is taking place.

CO-CHAIR CELLA: Thank you. Jack, and then Joe.

MEMBER NEEDLEMAN: Yes, to follow-up a little bit on Larry Glance's comment and also your statement about the use, so the intended use of this measure is looking within an institution

over time and not doing across hospital comparison.

MEASURE DEVELOPER: That's correct.

MEMBER NEEDLEMAN: Okay. And then I want to sure I understand the method of calculation here. So I've got a patient, within my 22 patients, I've got a patient on two meds. And there's been a discrepancy in one. So that patient gets a score on this measure of 0.5. I've got a patient on ten meds, and there's a discrepancy in one. And they get a score of 0.1.

MEASURE DEVELOPER: Correct.

MEMBER NEEDLEMAN: And then those scores are averaged across the patients to get the average score for the facility for that month.

MEASURE DEVELOPER: Correct, that's exactly right.

MEMBER NEEDLEMAN: So there is no adjustment for the fact that it is more likely to have an error the more complex the medication

regime of the patient is, that the patient on ten meds is more likely to wind up with a discrepancy because it's harder to reconcile than the patient on two meds.

MEASURE DEVELOPER: Right. And it was a linear effect, that the opportunity per med is the same if there is an exponential or a, you know, a quadratic, you know, then no, I don't, right, so ---

(Simultaneous speaking.)

MEASURE DEVELOPER: It'd be more linear. It assumed a linear relationship, you know, opportunities for error.

MEMBER NEEDLEMAN: And as you said, as long as it's within hospital measure and the case mix, and therefore the medication complexity in the patient stays reasonably constant over time, that lack of adjustment may not produce the perfect measure, but it's probably acceptable.

But you ought to be at least exploring in your data whether or not you see more discrepancies among the more medication-complex

patients and think about what that means for the measure you're using down the road.

MEASURE DEVELOPER: Yes, I agree. Yes, I mean, it could. Some patients' regimens are more complicated than others. We've written a few papers on the kinds of medication classes people are on, which ones are most commonly, you know, relating to discrepancies, the kind of diseases people have. And, you know, there may be an exponential effect, where it's exponentially harder. The patient's on many regimens.

But, you know, within a hospital, we assume, over the course of a few years, case mix is roughly the same within the limits of sampling. And then you are measuring something real.

You know, and the other point that I would make is a point that, you know, was made at the very beginning of this, which is that we don't really have another measure that actually measures quality.

Everything else is a lot of, you know, did you check a box, did you do a process. And we know full well that not only is that not an accurate measure, but it actually can be counter productive. Because hospitals will encourage their providers to check a box saying I did med rec, when that provider knows full well that they didn't actually have the opportunity to take a good medication history.

So instead of having transparency, instead of saying I couldn't take a good medication history, you're going to have to do one tomorrow, you know, they checked the box that said they're being measured. And so that just becomes counter productive. You know, this is the only measure out there that really actually measures quality.

CO-CHAIR CELLA: Well, thank you.

MEASURE DEVELOPER: And it's really been a powerful motivation.

CO-CHAIR CELLA: Thank you, thank you.

MEASURE DEVELOPER: Thanks again for

letting me speak.

CO-CHAIR CELLA: All right. I'm sorry to break in. We're a little bit out of order, and I apologize for that. We're still in the period now where only the subgroup members are supposed to be commenting. So Zhenqiu, you had your card up, if you still want to say something, and then Terri, and then Zhenqiu, then Sherrie, and then Terri.

Remember, you have to be in the subgroup at this point to speak. Everyone will have a chance to say things later, but I'm also going to ask If we try to be brief with what we have to say. Because we've got a lot to go through. And if do it at this rate, we'll be out Wednesday --

(Laughter.)

CO-CHAIR CELLA: -- by 5 o'clock.

MEMBER LIN: So I just want to clarify. I mean, this is proposed as a quality improvement measure for maybe past two years, right. Because otherwise, if you propose a

sampling scale of 25 patients, I get to Karen's point, no, Sherrie's point. If you have 25 complicated patients, and 25 simple patients, then you compare the difference across hospitals, right. That would be, you know, it just may not be fair.

And also as part of the evaluation, the agreement between two trained pharmacists, 77 percent, so obviously pharmacy, different hospitals use different pharmacies. And that's going to introduce a lot of potential difference, right. So how do we know? If you use this measure to be better between hospital difference, how can we be certain that a lot of it is not due to pharmacies different?

CO-CHAIR CELLA: All right, Sherrie.

MEMBER KAPLAN: So I just follow-up on Zhenqiu's and Larry's point that the precision of the estimate for a patient who's on two medications and one has the discrepancy is the same as the score for a patient on ten medications for which there are five

discrepancies.

And to me, that kind of risk adjustment is really called for, and maybe a recommendation. I don't know if it's true for this round or not then, but maybe a recommendation going forward to the standing committee for what is it, a call for risk adjustment?

Because it seems to me like that really causes precision of the estimate problems at patient levels. So if you're at that twice, and you strike out once, you're into saying that if you're at that ten times, then you strike out five times. At least I get that with scores analogies.

(Laughter.)

CO-CHAIR CELLA: It is the season for that. Terri?

MEMBER WARHOLAK: Well, I had some similar comments and concerns. One of the things that I was worried about, I don't remember when I was reviewing this that this was supposed to be a

within hospital measure only. And so it didn't seem clear to me reading that.

And so I know that a lot of measures are used inappropriately. So I wondered what's going to be done or what is done to ensure that it's used appropriately. And how is Leapfrog using it? Are they using it to compare between hospitals or within?

MR. LYZENGA: Maybe also to clarify that endorsement does provide a measure suitable for accountability purposes.

(Simultaneous speaking.)

CO-CHAIR CELLA: Okay, can we give the developer a chance to respond if you like, and then Karen.

MEASURE DEVELOPER: Back to you then for the--

CO-CHAIR CELLA: Yes, if you'd like to respond to the last couple of comments, please do.

MEASURE DEVELOPER: Absolutely. Leapfrog is definitely in agreement with me that

it is, it should be used to compare within the hospital and not between hospitals. And the measure developers at Leapfrog have been very supportive of that piece.

Yes, I think the comments made about precision are correct, you know. And so if the feeling of this group is that further risk adjustment is necessary, then we would obviously work with them on that piece.

Was there another comment that ---

CO-CHAIR CELLA: I think that's good. Karen?

MS. JOHNSON: Yes. I was just going to reiterate ---

MEASURE DEVELOPER: Okay, thank you.

MEMBER KAPLAN: Yes. I was just going to reiterate what Andrew said. Measures that are endorsed by NQF should be suitable for both QI and accountability. So it needs to be both.

CO-CHAIR CELLA: Jack?

MEMBER NEEDLEMAN: Okay. So I'm going to go out for some summary comments, or I am,

right, as a member of the committee. I had basically passed this measure when I rated it. I have some concerns about subsequently what the training materials are, but I'm not competent to evaluate that.

So I'd like the standing committee to look at that. Because that affects the usability of the measure and, to some extent, the reliability of the measure in a way that I can't assess either, even if I had the materials, that I can't assess.

So I'm still prepared to pass the measure on to the standing committee. Or I'd like the standing committee to look both at the, think about substantively the area level of agreement or disagreement among the two folks that are, you know, looking at it together, and whether we have a reliable enough measure in practice to be usable.

And likewise for the training materials, I'm thinking about whether or not the measure, which is inherently about the training,

is both usable and consistent with the pace with the adjustment.

CO-CHAIR CELLA: If I could just maybe ask you, Karen, to correct me if I get this wrong and then elaborate. You made a comment, Jack, about whether this passes on to the standing committee. Everything is going to go on to the standing committee, correct?

MEMBER NEEDLEMAN: Right.

MS. JOHNSON: For potential discussion, not necessarily for a re-vote.

CO-CHAIR CELLA: Right, right. So whether this passes this committee or not, the standing committee will see it. And so if you could elaborate a little bit, Karen, on what would be the difference in what they see if it passes versus fails. Because that may affect some peoples' votes.

MS. JOHNSON: Yes, so what we are going to do is every measure that comes out of the Methods Panel is going to go to the standing committee. Now, the standing committee, if they

so choose, for measures that do not pass, the standing committee can pull those for discussion.

That means they could have the same kind of discussion that you have here, but it doesn't necessarily mean that we will allow them to re-vote, okay. So some measures will be eligible for a re-vote. And those are measures where they don't have certain things completely wrong with them, right.

So that would be things like the methodology just didn't work at all, or the methodology was inappropriately applied, or they didn't meet our minimum requirements for endorsement, those kinds of things.

So what you would have to do for today is just vote. Do you think that this measure, we're voting only on validity, do you feel like these measures or this measure actually is a valid, can you, do you feel that this measure can reflect the correct conclusions about quality, okay? That's basically what you are trying to get to with your vote on validity.

I don't know if that helped or not. But not all measures will necessarily be up for a re-vote. And that will have to be, the way we're going to do it is we're going to figure out which ones eligible for re-vote or not.

And to be honest with you, I think we need to do that in consultation with our co-chairs. So that's not something that we can tell you today, but they will be able to pull and discuss but potentially not re-vote.

CO-CHAIR CELLA: Is there, I guess, follow-up questions, sorry for this, but is there such a thing, maybe not, in this name, but such a thing as a pass with caution or pass with notes?

MS. JOHNSON: There's always going to be no votes, right. So we don't give them just the pass, no pass. We give them a pretty extensive discussion point of what you guys are finding.

CO-CHAIR CELLA: So concerns will be sent --

MS. JOHNSON: They will.

CO-CHAIR CELLA: -- along with --

MS. JOHNSON: As voted on.

CO-CHAIR CELLA: -- any passed.

MS. JOHNSON: Yes. If you pass it, they will still see the concerns. If you do not pass it, they will still see the concerns. We will have to decide, at a future point, whether it's even eligible for re-vote. And the requirement that it has to be suitable for both accountability and for internal QI, is probably really the sticking point in terms of passing our minimum requirements.

CO-CHAIR CELLA: Okay, any other, thank you, any other subgroup comments? Let's open it up then. Any comments from anyone? Gene?

MEMBER NUCCIO: Yes, I have a couple of questions. You described the issue as quote, "unintentional discrepancies." Why did you choose the word unintentional? A discrepancy is a discrepancy. And you're not trying to determine whether or not the pharmacist made the

problem of a wrong medication done on purpose. So it's simply discrepancies, not unintentional, from my perspective.

The second question, you cite an example in Section 1B.1 of your MIF form of a rate of discrepancies. And this, I think, is a question that Larry and Sherrie have talked about. Did you choose not to use a percentage of discrepancies per patient and then average those percentages which would take into account whether or not at least the number of medications is taken into account?

Complexity is a different issue. But regarding complexity, were your pharmacists directed to look at simply inappropriate drug interactions, Drug A should not be used with Drug B, or if they were also looking at dosage discrepancies? And I could not find that in my quick reading of the materials. So those are my three questions.

MEASURE DEVELOPER: Can I answer?

MEMBER NUCCIO: Please.

CO-CHAIR CELLA: Yes, please.

MEASURE DEVELOPER: So I could, but before I do that, I just wanted to get back to an early point that was raised.

With the help of Leapfrog, we have greatly improved our training materials. We feel they are, you know, at this point people can really achieve reliable and valid, even full standard medication with which to measure discrepancies.

And we'd be happy to share those with the full committee so that, you know, folks can weigh whether or not they agree with us, that the training materials really are that good.

In terms of accountability, I hear what you are saying. I don't know if it helps alleviate that concern, but for example, our problem is, where we have seven hospitals, what we would do is say that every hospital has to either achieve a certain degree of improvement or achieve a certain threshold of excellence.

And by combining those two, I think

hospitals that, you know, have very different case mix can still both be held accountable in an absolute sense while, you know, depending on improvement within their own hospital.

To get these two specific questions though, by giving discrepancies per medication we are essentially doing a percentage. In fact, if you take that number and divide it by two, it basically is a percentage of opportunities for error, for which there is a natural error.

So in other words, you know, one error, one discrepancy for two meds, which could be considered a 0.5, is really a 0.25 percentage error rate. Because there's two opportunities for error for each of those two meds. Four opportunities for error if you one, so you have a one out of four discrepancy, you know, sort of error rate as a percentage.

Why we use the term unintentional, just to make it clear that, you know, providers make changes to regimens all the time depending on what people came in on. And we just wanted to

make it clear that, you know, adding an antibiotic within that regimen is not a discrepancy.

So if you were at the camp that a discrepancy automatically incorporates the term unintentional, then yes, it's redundant. But we did put it there to make it clear. But yes, it's an unintentional discrepancy. So if you just want to call that a discrepancy, that's fine. But changes are made. And we wanted to make it clear that it's only the ones that were not intended.

In terms of the types of discrepancies, so dose frequency formulation, all of those are considered discrepancies. You don't get counted more than once per order. It's an order with a discrepancy. Because an omission is just as bad or worse than both the dose and/or frequency discrepancy. You wouldn't want to count the dose and discrepancy twice and an omission only once.

So once an order is wrong, the order

is wrong. They account for the discrepancy. So hope that explains that. But all the different types of discrepancies are looked for as part of this process. So if there are any follow-up questions, I'd be happy to answer them.

CO-CHAIR CELLA: Thank you.

MEASURE DEVELOPER: Thank you again.

CO-CHAIR CELLA: Okay, Joe you've been very patient. You're next and then Sherrie. And again, let's try to be very concise.

MEMBER HYDER: The first question I had has to do with the type of measures. I think you have a definition of outcome as health state of a patient or change in health state. And I'm curious how this is classified as an outcome measure. And then I have a second question.

CO-CHAIR CELLA: Okay, quick answer on that one?

MR. LYZENGA: Just an answer that typically our developers indicate whether it's an outcome, or a process, or whatever when they do the submission. Sometimes we talk with them and

do a correction. I think in this instance we allowed it to come through as an outcome, although that is certainly arguable. Again, it doesn't quite fit in our usual categories. It seems to me a little bit more like something like an immediate outcome.

CO-CHAIR CELLA: Yes. I don't think we need to worry about that here. Let's go ahead, Joe, what's your next question?

MEMBER HYDER: The second question has to do with the gold standard. So I'm thrilled pharmacists are the gold standard. The way it's constructed though, pharmacists aren't the gold standard. It's the brief medical record review.

So the impression is that the medical record is the holy book containing all knowledge about what medication is required for the, what medication was relative to the patient's admission and the patient should be started on. In practical terms, that's never the case.

So I don't know that there is a gold standard in this case. And using either the

pharmacist or, specifically, the medical record review would cause me to question the validity significantly.

CO-CHAIR CELLA: Quick comment?

MEASURE DEVELOPER: Yes. So what these pharmacists do is a pretty regular process, interviewing the patient, getting corroborating sources, talking to BCDs, calling up the pharmacy, talking to family members, getting pill bottles.

They put all of that together, then they put together what they think the patient should have been on and what they're actually taking. So it's a pretty regulated process. It is not just trusting the charts. But in the end, what they come up with is as close as you can get to the truth.

MEMBER HYDER: So that's not specified in the measure as written, that has their payment strategy. And then on top of that, it doesn't allow for the physician directing care to say I'm going to withhold the medication in anticipation

of something two days from now. So that maybe the ascertaining strategy for that may be specify in more detail.

CO-CHAIR CELLA: Good point. Okay, Sherrie?

MEMBER KAPLAN: I'm just a little confused about the scoring now. Because, like, you said something like you're treating your role like as odds ratio, like the probability of being wrong and maybe because you have two outcomes. You either were discrepant or you weren't. And then so each opportunity has two possible outcomes, yes or no. Is that right? Are you doing it more like an odds ratio?

MEASURE DEVELOPER: It's not an odds that you take the number of discrepancies per medication for a patient.

MEMBER KAPLAN: No, but each medication has an opportunity to either be discrepant or not be discrepant. And that's taken into account in the scoring which is more like you do an odds ratio than a relative risk

assessment, correct?

MEASURE DEVELOPER: Yes. I mean, I guess if you take the discrepancies per med and divide it in half, you will get the number of errors per opportunity for error. We still present that as a numerical number and not an odds ratio. It could be ---

MEMBER KAPLAN: It's not?

MEASURE DEVELOPER: -- I guess a ---

(Simultaneous speaking.)

MEASURE DEVELOPER: You know, it would be another, like, X over N would be another way to analyze that data. But, you know, happy to talk about that more offline.

CO-CHAIR CELLA: Okay, thanks. Joe, maybe a final point so we can vote?

MEMBER HYDER: Sorry.

CO-CHAIR CELLA: Other Joe, Joe Kunisch.

MEMBER KUNISCH: Just real quick, the way I understand it, there's no exclusions here. In looking at the gold standard medication

history, it says shortly after admission, for example, within 12 to 24 hours.

So if there are no exclusions and you have, say, a homeless patient that comes in showing symptoms of stroke, confused, can't even give you their right name, there is no other information on them, how would that patient be treated if they're not excluded?

If you did everything to get that medication history, which really would mean you couldn't find anything, would you get a pass on everything? Or would you fail on everything, because you didn't get any medication history?

MEASURE DEVELOPER: That's a great question. If it's typically impossible to determine, with any sense of assurance, that you can figure out what they are on, I guess you would have to exclude that patient. You know, under most circumstances, there is a pharmacy, there's a medical record, there's something. If there's nothing then, sure, the patient would be excluded.

I think the way we say it in our application is that able to take a medication history, you know, after they are admitted to the hospital. If it's physically impossible to take any kind if history at all, then they would be excluded.

CO-CHAIR CELLA: Okay. Thank you very much. That was a good discussion. Good then, let's run the vote, Poll Everywhere if you're in the subgroup.

MS. WILBON: So Dave, sorry, so, Dave I think you're not eligible to vote. But everyone else in the subgroup, Sherrie, Zhenqiu, Jack, Jen, Christie, and Terri should be voting via the Poll Everywhere.

Everyone else, please pull up the SurveyMonkey at the link that was emailed out this morning. And we'll ask that you select Measure 2456 from the dropdown. And you do not have to enter your name if you don't want to. And let us know what your shadow vote is.

CO-CHAIR CELLA: Yes, who's that?

Let's continue on ahead. You all got the vote instructions?

MS. OGUNGBEMI: Yes. We're voting on the validity of Measure 2456, the overall rating on validity, taking into account the results and scope of all testing and analysis of potential threats. Voting is open. Your options are high, moderate, low, and insufficient.

CO-CHAIR CELLA: Are we able to move to 1623 while people vote, or do we need to wait? Do you want to do the break now?

(Simultaneous speaking.)

CO-CHAIR CELLA: So let's try to get through 1623 while you're voting. Karen, you're doing the slide. And then if you do need to take a bio break before the break, which is going to come after this discussion, please feel free to get up and do that. But let's try to get through this and catch up a little.

Go ahead, Karen.

MS. JOHNSON: You want to do that before ---

CO-CHAIR CELLA: Can we? Is it possible to put that screen up. Or do we need to wait for the voting to close?

MS. OGUNGBEMI: I can do both at the same time. It's there.

CO-CHAIR CELLA: Oh, okay. I'll just look over here. Yes.

(Simultaneous speaking.)

MEASURE DEVELOPER: I'm going to get off the line then. This is Dr. Schnipper. Thank you for letting me --

CO-CHAIR CELLA: Yes, I'm sorry. Thank you, Dr. Schnipper, from everyone.

MS. OGUNGBEMI: Has everyone in the subgroup voted that was eligible to vote? Yes, okay.

MS. JOHNSON: Let's go ahead. Do you have those results? We need to get the actual results of the vote into the transcript.

CO-CHAIR CELLA: Oh, okay.

MS. JOHNSON: Yes.

CO-CHAIR CELLA: Well, then we do

wait. All right.

MS. JOHNSON: Yes.

MS. OGUNGBEMI: So we have zero votes high, 50 percent moderate, 33 percent low, and 17 percent insufficient. So we're still at consensus not reached.

CO-CHAIR CELLA: Not reached, okay.

MS. JOHNSON: Okay. All right, our co-chair's wanting us to keep going, not have a break. So let's do Bereaved Family Survey.

PARTICIPANT: Karen?

MS. JOHNSON: Yes?

PARTICIPANT: I got that measure, and I'm not on the list --

MS. JOHNSON: Yes, you were on the list, for which one, 1623 or --

PARTICIPANT: 1623, but I'm not listed here.

MS. WILBON: Are you listed under subgroup voting eligibility? Is that where you're looking at? Okay, we only listed people who have special circumstances for voting.

PARTICIPANT: Oh.

CO-CHAIR CELLA: Yes, you're okay.

MS. WILBON: So you're good.

PARTICIPANT: I'm good?

MS. WILBON: Yes.

MS. JOHNSON: Okay. So Bereaved Family Survey, this is a maintenance measure. And it was previously reviewed by the Methods Panel and did not pass at that time. And unfortunately, that made a little bit of confusion for our group. And it was our fault, so we will take credit for messing up on that one. And I'll explain where that is.

So this is a measure that comes out of the Bereaved Family Survey. It calculates a portion of veteran decedents' family members who rate overall satisfaction with the end-of-life care in an in-patient setting as excellent versus all the other options, very good, good, fair, or poor.

It is based on one item in the Bereaved Family Survey, Item Number 18, overall,

how would you rate the care that the decedent received in the last month of life?

It is, of course, a PRO-PM. And it is calculated by taking the number of respondents that choose excellent versus everything else as long as there are at least 12 or more valid responses on all the items in the survey.

The level of analysis is a facility, and it is risk-adjusted with five factors. Just a couple of things we wanted to point out in terms of the initial analyses of this.

There were a few commenters or SMP folks, and this is probably people who are new, who were pointing out that some of the things in 1B and 3 and 4, those items, weren't filled out completely.

That's okay. This is not filled out completely, because those are actually due a little bit later in time. So really, you were only supposed to be looking at the specifications and the testing attachment.

So I can talk to you offline if that's

still a little confusing. But we don't have to worry about them not filling out the feasibility section. That's okay at this point.

There was also a note about waste of respondent data. So I think a little bit of confusion, there's this whole survey, lots of different opportunities for PRO-PMs that this is only being based on one item, that satisfaction item.

That also is okay. NQF is not saying that you have to use the entire survey. So developers can decide which measures that they develop and which ones they put forward to NQF for endorsement. So the fact is, they only put one PRO-PM. There could be multiple ones out of the survey. That's okay.

And then finally, the previous evaluation by the Methods Panel, we wanted to give you the previous testing attachment, more for your information only, so that if you were a previous evaluator of this measure, you could go back, and you can compare what was done before

versus what's done afterwards, if you were curious about that.

Unfortunately, we did not make it as clear as we should have which was the most recent testing attachment. And what we think happened is a few of you probably accidently looked at the previous attachment, not the most current one, okay.

So especially the last time around, there was concern about the reliability testing of the one item in the instrument. The last time around it was not done. This time, they did do tests, re-tests, and analysis of that one item. So that was done.

And there are a couple other things throughout there that makes us think, again, that maybe were looking at the wrong testing attachment. So we just wanted to make sure that we point that out, make sure that you understand, number one, that they did provide a different and more updated testing attachment. And they also, in their additional responses, hit those high

points that were mentioned in the evaluation.

So all that said, I'm not going to go through all the stuff on reliability, because the actual ratings for reliability, it passes with a high rating. So I'm not going to go through that.

But it did not pass under validity. And I think it could have been because, again, and we can talk about all of this, some of the main questions really was, was there any kind of testing done beyond just showing the rates over time, okay.

So let me stop there. We do want to give it back to the lead discussant, but before we do that, I just want to make sure, Dawn, are you on the phone or anyone else from the VA, the developer?

MEASURE DEVELOPER: Yes, myself and then two other members of our team, Josh and Anne.

MS. JOHNSON: Great, thank you, Dawn. So what we're going to do is we're going to have

our lead discussants just point a couple of things that they thought were the key issues. And we'll figure out were those the issues just simply because of looking at the wrong form, or were there additional things that we need to talk about? And then we'll go from there, okay?

CO-CHAIR CELLA: Christie?

MEASURE DEVELOPER: And do the panel members, team members have our response that we sent last week as well?

MS. JOHNSON: Yes, they do.

CO-CHAIR CELLA: Yes.

MEASURE DEVELOPER: Okay.

MS. JOHNSON: Yes. So our lead discussant is Christie.

MEMBER TEIGLAND: Yes, I'll go through a few things that I had noted on my review. One is that they did not adjust for the type of facility. And while they did note that there were differences depending on the type of facility, that they had risk adjusted for that. So that was a confusion, a point of confusion.

And I'm not clear why they didn't adjust for type of facility. That would seem to be important.

Just a small thing, they called the Elixhauser score, the co-morbidity score, a social risk factor. Which I think, I see some heads shaking now, I would question that.

They also noted, regarding the missing data, that there were differences. They did see differences in scores after they waited for response rates at the facilities. But the non-responses were not really addressed. And I think that could introduce some potential bias in this measure.

They also noted that facilities that, you know, that were in urban areas in the northwest or the Midwest, had the biggest improvements after risk adjustment and that those in rural areas, the mountain areas, the west, had the biggest declines in scores after risk adjustment. But they didn't use regions as a risk adjustment factor. So I'd like to understand a little bit more about why that

wasn't considered as a risk adjustment factor.

Yes, the face validity looks pretty good in terms of they did pick some facility process measures that they correlated then with that item. And they found good correlation that facility scores were higher when they seemed to do the right thing, like getting palliative care in the last 90 days of life and so forth, so that looked good.

My last comment is one that we presumably can't score them down on, but I need to comment on it. And there was no risk adjustment for social risk factors even though, again, they saw a disparities in how patients responded to the overall care, end of life care, based on racial, race categories.

So what they say is that that will let facilities off the hook. And what I would point out here is this is completely inconsistent with all of the AHRQ past measures which I adjust for race, ethnicity, income, education.

Because particularly, with these types

of surveys, cultural differences in the way that you'll respond, based on not only race ethnicity but income levels, people with higher incomes are going to actually tend to think their care is worse, right, than people in with lower incomes. So I strongly question that decision, not to even look at those characteristics. Yes, so those are some of my concerns.

CO-CHAIR CELLA: Let's have a, let's see, you've raised a few things that maybe we should give the developer from the VA a chance to respond to and then we'll have the subgroup ---

MS. JOHNSON: Dawn, do you want to take a couple of minutes to just respond to Christie's first set of concerns?

CO-CHAIR CELLA: Or we can move on if you prefer, and you can respond to everything.

MEASURE DEVELOPER: Good.

CO-CHAIR CELLA: Go ahead.

MEASURE DEVELOPER: I mean, Anne and Josh, would you say that the responses that we gave in our response last week are, do you want

to add anything in addition to those?

MEASURE DEVELOPER: No. This is Josh, not particularly. I guess in looking backwards, I mean, those are really, each of those are things that we considered. And we still struggle with them, in particular, you know, what to adjust when it comes social factors.

From a VA perspective, we're accountable for the patients that we get, the veterans that come to us. And so, again, from an accountability standpoint from within the VA, we don't adjust a way for having patient populations that face certain challenges.

You know, our goal is to provide the same level of care and to hold all of our facilities accountable, regardless of the social factors. What's different than--, why don't we adjust more for medically sound patients, that's a different story.

But that's fine. That's kind of our explanation, and that comes from the VA philosophy more generally. At least that was one

of the points that was raised. That remind me, I should have written the check list, the list down. Others, Dawn or Anne?

MEASURE DEVELOPER: Also the one item was, we do our risk adjustment, we adjust for facility complexities which is the high/low complexity. And that was the question regarding the differences in the facilities.

MEMBER TEIGLAND: That's not listed as one of the five though. And so it's not clear then in the, you're not listing it as risk adjustment, checking adjustment, as far as I can see. There are five. And that was not one of them.

It was, age, number of co-morbidities, Elixhauser, primary diagnosis on last admission, relationship with next of kin, and the mode of the survey. There was nothing about complexity of the facility.

MEASURE DEVELOPER: Okay.

CO-CHAIR CELLA: So we're going to continue around. And then you'll have another

chance to make comments. But let's go to Jack, subgroup members for now.

MEMBER NEEDLEMAN: Okay. So this is one of the measures I found in low level of validity on. And I will, you know, leave it to, you know -- And there were several reasons for that.

So, one of them is the intent of the measure, what it's supposed to be measuring, is not clear. So, it's very hard to measure whether it's valid or not.

The clearest statement of the intent is in response to Issue 2, which is given the alignment of patient family preferences with treatment is the cornerstone of optimal care.

The purpose is to assess family's perceptions of the overall quality that is received. Basically, did we get what we wanted for our family member or friend?

And that's okay. Fine. But if that's the face validity, if that, the question asks that, and they get an answer to that. So, on

face validity bases it's fine.

The risk adjustment model is a mess. There's, you know, they include the facility complexity. They don't include social risk factors, which we know affect responses for surveys like this.

They do include the disease complexity when it's not clear to me that, and they have not asserted that it's harder to provide care the family expects for a more complex patient. And therefore, we have adjusted for it.

Unless you believe that there's no reason for the patient complexity to be in the risk adjustment. But if they know that it's harder to do that, then they ought to say that. And they haven't.

I've got no problem with the face validity. I've got a lot of problems with the construct validity assessment. And they responded to that in their comments. And I'm not convinced.

You've got a beta coefficient of .03

on a measure that varies by .5 or .15. Anyway, it's a very small proportion of the variance that's explained. If we had an R square, or an R for the correlation it would be very low.

And the argument that because we've got a large enough sample to make a very low explanation of variance be statistically significant is not saying you've demonstrated to me that you've actually got a correlation here, which is what construct validity is supposed to be looking at.

Given that some families want high levels of intervention at the end of life, and others want low levels of intervention, it's not surprising to me there's a low correlation between whether you prove to satisfaction whether you got high or low.

But, and therefore, I don't believe the construct validity measures are terribly important here. I'd go with the face validity. But I don't want the developers to believe that they have demonstrated construct validity with

correlations and variance explained that's this low.

CO-CHAIR CELLA: Thank you.

MEASURE DEVELOPER: Can I respond to that as a developer? Or do I have to --

CO-CHAIR CELLA: Yes. No, go ahead.

MEASURE DEVELOPER: I would argue that in fact the point of construct validity is not in the effect size, which is something that it seems that there's a primary concern.

It's in the direction of the effect, and whether that direction is, we can rule out is that direction driven statistically by chance alone.

So, I think in fact we do establish construct validity, in that the direction of effect is, follows theoretically plausible directions. And it's statistically, it rules our chance as the key driver behind that.

Now, whether it's a huge effect or not, that's a, to me a different question than getting at, you know, what I think of as

construct validity.

You know, we approached it, our, the nature of these linkages in directions that we might expect based on theory and prior empirical. And that was the goal. And not necessarily that these are the, you know, we're getting R squares of .8.

And these are defined as exploring the great deal of the variability. It's more, if you believe these things are appropriately related in construct validity as established, you would expect a positive relationship, and ruling out chance as a cost.

CO-CHAIR CELLA: Thank you.

MEASURE DEVELOPER: And that was our thinking behind it.

CO-CHAIR CELLA: Thank you. Sherrie.

MEMBER KAPLAN: Yes. Following up on that I just, I have another concern. But what constitutes a clinically meaningful difference?

Statistically significant differences can be had for a variety of reasons. It does not

always need to be clinically meaningful. And so, small differences between facilities by two to three percent in the between facility process measures, with and without failing of services, for example, to me isn't a real compelling argument, even though the direction is correct.

So, I'm a little bit push back on this kind of call of, as long as it's statistically significant and positive, which is good.

My concerns are two. One is, the magnitude of the response rate varies so substantially by facility, between 29 and 73 percent, with about an average of 48 percent response rate.

And then, when you look at that, waiting for the non-response rates changed the facility score by more, well, about a quarter of a standard deviation for about a third of the facilities.

So if, that's kind of a really important issue. And it brings up the issue of selection bias as a potential concern, given the

patient characteristics between non responders and responders. So, that's question number one.

And then, question number two. On Figure 2, when you looked at the facility level variation in the bereavement survey, what I would have like to have seen was some standard error bars on that.

Because that looks like, between facility differences, okay. You can't really look at between facility differences if you looked at the standard error bars.

But one could argue that you could look at the top and the bottom quartile, for example. And those would be substantially enough different from each other.

So, you could make the case for using it, rather than discriminating individual facilities one from another, discriminating the top and the bottom quartile. So, could you kind of address those two issues? The issue of the real missing data, you know, on response issue.

And then, the issue of this, could you

put standard error bars on Figure 2, and give us some evidence that there's some compelling differences here? At least interpretable differences between quartiles?

MEASURE DEVELOPER: Yes. I don't think standard error bars are an issue that we, you -- Yes, we could add that to that table. And then, the non response bias adjustment, and the selection bias. Josh, would you be able to respond to that?

MEASURE DEVELOPER: Well, it, yes. The reason, I think what we're, we're always facing the potential for selection bias due to non-response in, you know, in all of our work.

We do a non response adjustment. And we, you know, it's only as good as the variables that you have in your dataset that might be driving non response, the selection.

In this case we happen to have a very rich dataset, because we know a lot about the veterans before. So, you know, we're not relying on just the basic demographics for our non

response adjustment.

But we have all of their clinical administrative data that I think that we can. And we could go to that non response adjustment. So, I think that we provide a pretty rich set of variables to, you know, I think.

Again, we can only adjust for what we measure here. But we have more measures than our typical for surveys of this kind, to adjust for potential bias and non-response.

CO-CHAIR CELLA: Thank you. Jack.

MEMBER NEEDLEMAN: Yes. I do not want to get into an extended argument with the developers about this. But I will simply say that as a voting member of the Committee, my attitude is, sign is not sufficient.

Magnitudes matter. That at least if you're trying to argue construct validity, I'm not looking for R squares of .8. I've considered construct validity demonstrated with Rs, where it's well below that. But here the R square is really low, as measured by the coefficient.

And frankly, all the measures of patient, and even sign is hard to interpret. All the measures of appropriate end of life care basically are built around doing less, when it's not going to have any effect on the patient, other than torturing them.

And I agree with them. But if you have in the VA a substantial number of families that wanted more intensive care, and they got it at the end of life, you would have the wrong side, as they've interpreted it, on this coefficient, because more was better. While the other measures of good care say less is better.

So, sign is not sufficient. I do think the magnitude of the, if you're going to argue correlated, correlation proves, validates the measure, you've got to have some correlation at an appropriate level. And this one doesn't.

So again, on the face validity I've got no problem with the measure. It's the other measures that are here, partly driven by the need by NQF to have some other measure beyond face

validity, that has led to the inclusion here of these measures.

I am still concerned. So that, I just wanted to get that on the record. And I do have concern about both the risk adjustment model, and how low response rates are being handled in the adjustment.

CO-CHAIR CELLA: Now, I think -- Okay.

MEASURE DEVELOPER: Can I get a follow-up on more specific, on what are your concerns specifically about the non response adjustments?

MEMBER NEEDLEMAN: Well, it's just a matter of how it's, you know, it introduces a substantial level of potential bias in the results. And you're handling it through a propensity model, which I'd like to, you know, I think is okay. But I'm not sure I've gotten enough data to fully assess it.

CO-CHAIR CELLA: Okay. Christine?

MEMBER TEIGLAND: Yes. I just want to make another quick comment about the social risk

adjustment. And taking your logic, you know, to the factors that you did adjust for, are you saying that you can excuse for age for old people? Because you did use that as a risk adjuster.

MEASURE DEVELOPER: I thought, would you --

MEMBER TEIGLAND: Age.

MEASURE DEVELOPER: Age? Sure.

MEMBER TEIGLAND: Yes. Why use age, but not race, ethnicity? Yet --

(Simultaneous speaking.)

MEASURE DEVELOPER: Yes. Age being a very good proxy for clinical complexity. It's not a perfect proxy. But age being a very big risk factor for --

MEMBER TEIGLAND: But do you have --

(Simultaneous speaking.)

MEASURE DEVELOPER: -- and that's the complexity.

MEMBER TEIGLAND: Do you have co morbidities for, and primary diagnosis?

MEMBER NEEDLEMAN: So, since you didn't say --

MEASURE DEVELOPER: Correct. And, right. And age in addition to those.

MEMBER NEEDLEMAN: Okay. You didn't say it in any of your documentation. Is your argument that it is inherently harder to satisfy people, families in end of life care when the patients are more clinically complex? Because that's what's implied in your risk adjustment model.

MEASURE DEVELOPER: I think we need to think about that. So, yes, I think --

MEASURE DEVELOPER: Yes. VA hospitals vary on the clinical complexities, based on the hospital.

MEMBER NEEDLEMAN: Yes.

MEASURE DEVELOPER: And it is harder to, yes, it can be more complicated. So it isn't, what you're arguing is correct. We are saying that implicitly we mean it explicitly.

They have, the more that is being

done, the more acute the care up to the point of death, is a potential risk factor for being less satisfied with end of life care, versus if it's a required hospice death.

CO-CHAIR CELLA: Okay. Thank you.

MEASURE DEVELOPER: We could make that more --

CO-CHAIR CELLA: Thank you.

MEASURE DEVELOPER: -- and actually back the --

CO-CHAIR CELLA: Thank you. We're going to open it up now to the full Committee. And I'm going to ask the developers, if there's more than one full Committee response or comment, please take notes, and have one response at the end.

So, let's go around and see how much we have from other people, and then have one response back from the developer. Okay? Go ahead.

MEMBER GLANCE: Quick process issue, Karen. When I try to bring up the source

documents on the web page, most of them give me an error, the PDF documents. So, it becomes very difficult to sort of like look, you know --

CO-CHAIR CELLA: Look at it while we're talking?

MEMBER GLANCE: Yes, exactly. So, that may be something we can fix. A couple of things. You know, I think when you're evaluating score level validity, risk adjustment really is central to doing that. Because you're looking at predictive validity.

It would be very difficult to look at construct validity. Because you're trying to compare one concept to another concept. You're not really sure how well, you know, if there's a really good argument, that they're correlated or not.

But at the end of the day if you can, if you have risk adjusted measure, and it has good predictive validity, meaning that the statistical model used for risk adjustment, if it's a logistic regression model, has good

discrimination, has good calibration, then you have a way, in a sense, of looking at, of predicting what the outcomes for a particular patient should be, conditional on their risk factors.

And then, if you have that for all the patients that a hospital treats, then you can compare the observed and the expected. And I'm really over simplifying things. Because there are different ways of doing this. It doesn't have to be an O to E ratio. It can be a P to E ratio.

But anyway, the point is, at the heart of score level validity is predictive validity. Okay? So here, what I'm not hearing very much about is, how well does this risk adjustment model perform?

And I couldn't get, it looks like the outcome is binary, right? So, I assume that this is a logistic regression model? What was the C statistic for the model? What was the HL statistic? What did the calibration graph look

like? Did they do this in a validation dataset, independent from the development dataset? That's my first question.

The second one is with regards to whether or not we should be adjusting for clinical complexity. That's really an empiric question. Whether or not this is important or not, you'll see it from the data. You'll see it from the risk adjustment model itself.

And then, the third comment I would make is, as we talk about reliability, or score level reliability today, I think there are lots of different ways of doing this. And something that we've talked about extensively at our meetings.

Okay. I'll look at it. But one of the ways is to look at the signal to noise ratio. Another one is to do split sample test, retest. A third measure, and I think that, you know, this is where it gets a little confusing, and I don't think it should be used to look at score level reliability.

Is when you construct a hierarchical model or a mixed effect model, and based on that model itself you calculate the intraclass correlation coefficient. And that's a very, it's very, very different from the ICC that we calculate when we do the split sample test retest.

And so, that ICC, the latter one where you're looking at just the intraclass correlation coefficient, based on the mixed effect model, it's not what we should be looking at, to look at reliability. Okay.

When we talk about the signal to noise ratio, based on hierarchical modeling, what we're doing is essentially we're looking for the signal portion. We're looking at the variance for all the provider effects.

And then, for each provider we're dividing it by the noise for that individual provider. And then, that gives us a signal to noise ratio for each provider. And then, over the entire sample we look at the median. And

that gives us the signal to noise ratio.

I mean, I don't mean to spend too much time. But I think this is really important. Because these issues are going to come up over and over again.

So, I guess my questions for the developers is, can you tell us a little bit about the performance of the risk adjustment model?

And can you tell us how you went about looking at reliability? Whether it was signal to noise, or was it split sample test retest approach? But how you did that, and what the results were.

MEMBER KAPLAN: Can I intervene? Because this is going to get us into a tangled discussion about what NQF is going towards, and where NQF is providing guidance now.

Because when they did that in the 18 submission, the 125.18 submission, the ICCs turned out very low reporting for --

MEMBER GLANCE: But how did they --

(Simultaneous speaking.)

MEMBER KAPLAN: As far as I can tell they did it the correct way, Larry. And if that --

MEMBER GLANCE: Did they do a signal to noise ratio?

MEMBER KAPLAN: They did. And when you include that extra denominator term the variance between facilities is not, versus within, across patients within facilities, it probably is, that's going to get us into a more complicated discussion about the magnitude of that ICC, the ruling set going forward is. Because --

MEMBER GLANCE: But, but --

MEMBER KAPLAN: When people have done it the right way with HCAHPS, it turns out exactly --

MEMBER GLANCE: But --

MEMBER KAPLAN: -- the same way. It's very low.

CO-CHAIR CELLA: Let's park that. We'll hope to have time at the end of the meeting

for --

MEMBER KAPLAN: But the risk adjustment thing --

CO-CHAIR CELLA: Let's --

MEMBER KAPLAN: -- is Larry's question.

CO-CHAIR CELLA: Let's keep the focus now on what developers of this measure need to answer, which, and you've asked three questions, Larry.

Are there any other comments from anyone on the Committee? Or we'll turn to the developer for their final comments. Okay. So, that's it. So, back to the VA developers. If you could respond to Larry's questions?

MEASURE DEVELOPER: I mean, we have published our risk adjustment method. And we provide these statistics, and the calibration, and things like that. So, I would just have to, I would just need some time to pull up that publication and send all that data.

MS. JOHNSON: I can tell you what it

is. And I don't know if it's too late for you guys to pull it up or not. See statistic --

(Simultaneous speaking.)

MS. JOHNSON: .535. They also did a EP information criterion, 36128.58. And for calibration they did Hosmer-Lemeshow. And got a P value of .1827.

CO-CHAIR CELLA: What's the sample size, Karen?

PARTICIPANT: Small.

MS. JOHNSON: Good question.

PARTICIPANT: Oh, wait a minute. You have 146? So, it would be a level of --

MS. JOHNSON: Sample size, I don't --

PARTICIPANT: You can figure two for--

MS. JOHNSON: Yes, 146.

PARTICIPANT: That's really small.

MS. JOHNSON: 146 facilities.

(Off microphone comment.)

CO-CHAIR CELLA: Any final comments from the developers?

MEASURE DEVELOPER: I do just want to

say in our case model for, I think there was a concern about the primary diagnoses that were associated with the system performance measure.

And then, in our prior work we have shown that, you know, that certain diagnoses, you know, that are associated with the performance measure.

And so, those are, that's the reason why we include that principle diagnosis in our case risk adjustment model, and why it's even more important when you're making facility level comparisons.

When you have higher proportion of patients such as with cancer, for example, and we know that cancer is associated with the performance measure, you're going to want to include that in your risk adjustment model. So, I just want to point that out.

CO-CHAIR CELLA: Thank you.

MS. JOHNSON: And, John, this --

MEASURE DEVELOPER: And just -- Oh, go ahead. Sorry.

MS. JOHNSON: Sorry. This is Karen. Just one more question. I think we still are a little confused. You did say that you do take into account the low versus high complexity.

We just don't know how you're doing it. Because it wasn't one of the five risk factors that you mentioned. And you said you did a non response adjustment. So, did you just mean to say you had six risk factors? Or are you doing it some different way?

MEASURE DEVELOPER: We just meant we just excluded it. But mistakenly excluded the variable complexity level in our original submission.

MS. JOHNSON: Okay. So, it should be six factors, not five?

MEASURE DEVELOPER: Correct.

(Off microphone comment.)

MEASURE DEVELOPER: And we also, with our statement that we submitted last week, we do say that we tried to follow other NQF endorsed surveys for choosing our risk factor variables.

And we included a reference in our response.

But the CAHPS, I guess that methodology, we tried to follow that as closely as possible with the variables that we had available to us.

And race, ethnicity, which is one of the social risk factors we all were very concerned about, was not included in that CAP adjustment methodology.

MEASURE DEVELOPER: Yes.

PARTICIPANT: Got it.

MEASURE DEVELOPER: Yes. And that's in direct, there's a reference that we included in our response. It's a direct link to AHRQ guidance on making adjustments to healthcare quality for us.

And so, we followed that guidance as closely as possible. And those were where they said, you know, include age. Co-morbidities, those are a good idea. But it's not considered appropriate in most circumstances for other socio-demographic characteristics, such as race

and ethnicity.

Because again, for the reasons that we laid out in our response, that would essentially bury information that could reveal what we might think is unacceptable disparities. So, we followed AHRQ's guidance on the--

MEMBER TEIGLAND: Can I respond to that? AHRQ uses race, income, education to risk adjust all of their CAHPS measures, which are survey based measures, which are listed. So, I disagree. I think you misinterpreted it.

MR. CARTER: Patrick, do you have a comment? Thank you, Christine.

MEMBER ROMANO: Yes. I was just going to point out that education is explicitly included in all the CAHPS risk adjustments. So, I'm not sure that race ethnicity is--

MEMBER TEIGLAND: It is.

MEMBER ROMANO: But certainly education is. And it's been well validated as a measure that is predictive of scores on these types of instruments. So, I don't know if that's

something that was outside the realm of the developers's success. But certainly --

MR. CARTER: Can I --

CO-CHAIR CELLA: Okay.

MEASURE DEVELOPER: Yes. Unfortunately education is not a valid variable as corrected in the VA in the patient's medical records.

CO-CHAIR CELLA: Okay. Thank you to the developers. I think we have no other comments or questions. Yetunde will set up the Poll Everywhere for the subgroup members to vote. And everyone else, SurveyMonkey.

And while we are doing this we are going to start our compressed ten minute break right now. So, because we're --

(Off microphone comment.)

CO-CHAIR CELLA: -- falling behind even more. And we're going to, we're determined to catch up. So take, make your vote, and take your break. And we'll be back in ten minutes.

MS. WILBON: So, I think we're

actually going to take our lunch early.

CO-CHAIR CELLA: Oh.

MS. WILBON: We're --

(Laughter.)

MS. WILBON: So --

CO-CHAIR CELLA: All right. I had a different plan to, I'm meeting with Sherrie and Dave, and we're going to -- Go ahead.

MS. WILBON: Oh, okay. Sorry if I'm --

CO-CHAIR CELLA: We have to eat now?

MS. WILBON: Well, they're getting lunch ready for us.

CO-CHAIR CELLA: Yes.

MS. WILBON: I think they'll be ready for us very shortly.

(Off microphone comment.)

MS. WILBON: If everyone has submitted their vote, do we read the results?

PARTICIPANT: I haven't even started yet.

MS. WILBON: Oh, okay. So, once we do

that, I think by the time we get everyone's votes, get the results on record, we'll release everyone for lunch.

We may need a couple of minutes, because we are a little bit early to give our support staff an opportunity to get the lunch ready. But let's take a 30 minute break for lunch when we do that.

We do also need to go out for public comment. And I do think there has been a couple of comments that have come in through the chat feature, that we want to address. I think by the time we do that we actually may be on track to go to lunch around 12:30 p.m.

PARTICIPANT: Yes.

MS. WILBON: So, if you guys would just bear with us a bit, I think for, you know, powering through the morning, we'll try to get ourselves back on track this afternoon. So --

CO-CHAIR CELLA: : So, what are we going to do between now and 12:30 p.m.?

MS. WILBON: We're going to vote.

We're going to do public comment.

CO-CHAIR CELLA: : Public comment's at 12:30 p.m. We can do it early?

MS. WILBON: Yes. We'll do it a little early.

MS. OGUNGBEMI: All right. So, we are now voting on 1623, the overall rating of validity. Your options are high, moderate, low, and insufficient. We are looking for six votes of the subgroup panel members, and the full everywhere. And the shadow vote can come from the SurveyMonkey.

MS. JOHNSON: So, subgroups are Sherrie, Zhenqiu, Jack, Jen, Christie, Terri. Everybody else will be SurveyMonkey.

MS. OGUNGBEMI: So, we have five participants.

(Off microphone comments.)

MS. OGUNGBEMI: So, we have zero votes high, 20 percent moderate, 80 percent low, and zero insufficient. So, the measure fails validity.

MS. WILBON: So, let's take this, hopefully everyone else has submitted their shadow vote via SurveyMonkey. Thank you for doing that.

And at this time we'll open it up for public comment to those in the room. Okay. Is there anyone on the phone who would like to make a comment? And we'll also be checking our chat box, to see if there's any other comments that have come in for the Committee to consider. Do we need to open up the lines?

MS. OGUNGBEMI: No. They're already open.

MS. WILBON: Okay. If you are on the phone, and you'd like to make a comment, please just unmute yourself. Or send a chat if you're not able to get through.

MEASURE DEVELOPER: Is the public comment open to everyone, including the developers? Or is that section over?

MS. WILBON: Sure. You can make a comment.

MEASURE DEVELOPER: Okay. I just wanted to make a basic comment that it seems, just based on the discussion on 1623 that we just had, that all of the panel members may not have seen the most updated testing attachment, or at least evaluated it thoroughly, in addition to our response to their concerns.

So, I guess I just, it just seemed like, from the conversation that we just had, that we've already answered those questions. And a lot of people weren't aware.

MS. JOHNSON: So, this is Karen from NQF. We did make sure that the panel knew and, yes, Christie is nodding her head. They did see the most recent one, and your comments. And as a matter of fact, in the discussion guide that we made available we have it in there as well. And --

MEASURE DEVELOPER: Okay.

CO-CHAIR CELLA: I think, just to say, I mean, as Chair, that discussion, I think what we heard was a difference of, was disagreement,

difference of opinion, and not lack of information.

MEASURE DEVELOPER: Okay.

MS. WILBON: Any other public comments?

(Off microphone comment.)

MS. JOHNSON: Okay. One quick question before we let you guys go for break. We didn't set up a formal dinner tonight. Would any of you, and if so, how many of you would be interested if we set up a dinner at a local restaurant?

Would you, is this something you'd be interested in going to? Or do you have your own plans for tonight, and really don't care what the rest of us --

MS. WILBON: If so, raise your hand --

MS. JOHNSON: Yes. Raise your hand --

MS. WILBON: -- so we can see how many people --

(Simultaneous speaking.)

MS. WILBON: -- to make a reservation

for.

MS. JOHNSON: One, two, three, four, five, six, seven, eight. Eight?

CO-CHAIR CELLA: Good group. Good crowd.

MS. JOHNSON: Okay. Okay. We will set up something tonight for those of you who want to go. We'll make it for eight, or ten, or 12. So, there will be a little room if somebody changes your mind. Thank you. Thirty minutes? Come back at --

MS. WILBON: Yes. Let's come back at --

MS. JOHNSON: 12:50 p.m.?

MS. WILBON: Ten til.

CO-CHAIR CELLA: Okay. Let's say a quarter til. So, you really, we really start at ten til.

MS. JOHNSON: We're shooting for quarter til. Yes.

CO-CHAIR CELLA: Sherrie, can you meet with Dave and me for just a sec?

MEMBER KAPLAN: Sure.

CO-CHAIR CELLA: Thanks.

(Whereupon, the above-entitled matter went off the record at 12:19 p.m. and resumed at 12:51 p.m.)

CO-CHAIR CELLA: So here's our catch-up strategy, which is going to -- we're going to succeed. Know it already. The measure developers, NCQA, are here joining in the discussion. We are talking -- we're going to talk about three measures: 0575, 0059, and 0061.

And we're not -- they all passed, so we're not necessarily going to revote here. It was voted forward for discussion. Sherrie's going to lead that -- or Jack will have something to say, but Sherrie was the person -- I'm sorry if I outed you, Sherrie. I wasn't supposed to. But wanted to discuss something and it probably does not have implications about the vote itself.

So, Sam is going to start, lead us off by talking about the three measures, give a little background, and then we'll hear from

Sherrie and Jack and then have a discussion.

MEMBER SIMON: Very good. Thanks, Dave. I'm going to use my outside voice so everybody can hear me, I hope.

So, just a couple of things to highlight. So Measure 0575, which you'll see here on the screen, comprehensive diabetes care.

Just want to briefly go over the measure descriptions. This is the percentage of patients ages 18 to 75, this page, with either diabetes Type 1 or Type 2 whose most recent HbA1C level is less than 8 during the measurement period.

So this measure incorporates a number of data sources, both claims using prescription claims or otherwise, HR data, electronic health records, as well as paper medical records.

So when this initially reviewed by the panel, we got a pass, as was mentioned, both for reliability and validity, which were ranked, on the aggregate level, as moderate. Next slide, please.

And you'll notice this companion measure was similar. So, this is 0059, comprehensive diabetes care, HbA1C poor control, which would be greater than 9. So this is simply a percentage of patients from the same age group category, 18 to 75, with diabetes Type 1 or 2 that were above 9 percent.

It follows a comparable pattern. In terms of the data sourcing, it's exactly the same. And it achieved moderate ratings on both reliability and validity in the aggregate voting.

And lastly, Measure 0061, comprehensive diabetes care blood pressure control, less than 140 over 90. Same patient categories, similar numerators and denominators in terms of the structure, but this time we're looking at blood pressure being less than 140 over 90. The reliability and validity rankings in aggregate were moderate and high, respectively.

I just wanted to mention that the reliability and validity testing was very

comparable between these three measures, but the reliability testing, they used the traditional methodology outlined by Adams with a signal-to-noise beta-binomial methodology. And for the validity testing we're looking at correlations between the measure of interest and four or five measures of a comparable quality domain using Pearson correlations between them.

I'll go ahead and hand it over to Sherrie, at this point, to discuss further. You're up, Sherrie.

MEMBER KAPLAN: So this is not for immediate voting purposes, just to reiterate, because the measure's already passed. This is a sort of opportunity, I'd think, for NCQA and NQF to kind of lead the way in maybe improving the way we're thinking about these kinds of measures.

In I believe it was -- or, I forget, 1990-something or other, NCQA had this provider recognition program going, and they looked at, and we looked at with them, 210 physicians and 7,400-odd patients and actually started to look

at what it would take to get a provider level estimate of quality for diabetes care.

As Daniel and I were discussing, all measures have measurement error. Glycemic control, I mean, hemoglobin A1C has measurement error. Blood pressure has a lot of measurement -- lipid level measurement error.

Well, when you start thinking about these kinds of measures at the patient level, most of them is not considered when you're sort of thinking about the dichotomy. You're either over or under the hemoglobin A1C value of 8 percent. You're either over or under the value of this.

But there's error associated with this. And when we dichotomize like that, we don't think about that category in the middle where there is measurement error. So we don't, beyond these thresholds, we don't put a confidence interval, for example, to take into account measurement error at this level.

Well, what NCQA and we were able to do

at that point, and this is like from 2000-and -- it was published in 2009, so I think I'm beyond the seven-year where they may have been a conflict of interest. But this is the kind of thing where, when thinking about these kinds of dichotomous measures, if they're all measuring the same thing, diabetes quality, this is an opportunity.

One of the ways to improve precision is to measure more things about the patient. We do that naturally when we do CAHPS surveys and other kinds of surveys. We add up different things that are supposed to be measuring the same concepts. This would be an opportunity to create a composite construct where you could actually -- because at the patient level this is dichotomous. It's either zero or one.

You can't really think about finding the partition variance associated with patient level of error because it's zero-one. But if you created an index, you're either zero-one on this measure. You're at zero-one on the lipid

measure. You're zero-one on the glycemic control measure.

They're all measuring diabetes quality. You could create a composite that then you could estimate the patient level variance on. So when you're comparing between --

(Simultaneous speaking.)

CO-CHAIR CELLA: Hello? Is someone trying to talk on the phone? Go ahead.

MEMBER KAPLAN: Anyway, so when you're comparing between plans on diabetes quality, you could actually look at the valid patient level variance and between-plan variance and get more of an estimate of how good the diabetes quality is across measures of what you think is the same thing.

So this little diatribe was more on the lines of going on from where we are in helping us consider how you'd consider working forward to look at plan performance with that standard error var around patient level variance within plans, you know, across these measures and

so on, and really think more broadly about can we improve precision for the purpose these measures are now being put to, which is discriminating between different plans?

So that was my bid. Did I make it, Dave?

CO-CHAIR CELLA: Great. Thank you.

MEMBER KAPLAN: For a moving forward argument.

CO-CHAIR CELLA: Jack, did you want to make a comment since you're a reviewer for the first one?

MEMBER NEEDLEMAN: Yeah, I actually rated this one insufficient, my first one. And it's not because I don't think the measure has a lot of positive things about it. There were two issues that I think NCQA needs to think about as they move measures forward.

One of them is they clearly have some method of treatment of missing data. It did not get clearly explained. My understanding of the NCQA process is the plans implement the measure

and send summary statistics up to the NCQA. That's different from CMS which is running it all through their claims databases so they have a lot of control over the way the analysis is done and the risk adjustment is easy.

So, missing data is supposedly dealt with. I would like to know how on the next set. I'd like to know how on this one, actually.

But the other thing is the risk adjustment. And they did not do any -- they did stratification by plan size. So the Medicaid plans are looked at together. The privately insured plans are looked at together. And there's a fair amount at SES discrimination and just by the type of plan, but I'm not sure it's enough.

And NCQA argued that site-to-site without any detail showed no variation and appropriate care within the plans by SES, and implied that meant no need for SES adjustment.

But the other argument for SES adjustment is that things outside of the control

of the providers are influencing outcomes. So, the lack of ability to exercise, or limitations in your neighborhood related to food availability, might well influence the ability of a patient with diabetes to control their diabetes. And that will be differential by neighborhood, not because of anything the providers are doing. Quite the contrary.

The usual argument against SES adjusting is you don't want to let the providers off the hook for poor performance with populations. But here the issue is there may be some elements of the neighborhood or family or work circumstances of patients in different SES status that influence the outcome that are outside, truly outside the control of the plan.

And so I am not completely convinced that that's the case as plans move to more population health initiatives. But I do think that the simple dismissing of the need for SES adjustment beyond the stratification by plan size is a little facile. It makes it easier for the

plans because NCQA doesn't have to give them a risk adjustment model to apply in their data analysis.

OPERATOR: I'm sorry. There's been an internal error. You will be disconnected now.

(Laughter.)

(Simultaneous speaking.)

CO-CHAIR CELLA: Are you really done there or --

(Laughter.)

MEMBER NEEDLEMAN: No, it's just that -- again, it's the simple issue of it feels to me a lack of risk adjustment in the NCQA measures are driven in part by the fact that the calculations are decentralized to the plans and they try to reduce the complications to the plans to do the analysis. And the rationality here feels a little facile, and I'm not sure it's right. So I'd like to see some more discussion of that with future NCQA measures, even though we passed these ones on this iteration.

CO-CHAIR CELLA: So let's go ahead to

Christie and then maybe give NCQA a chance to say something. Go ahead, Christie.

MEMBER TEIGLAND: I just had two quick comments on that. One is that these are actually medical record reviews, so I think up to 400 is the maximum number of patient records that they pull, because you're not going to find these HbA1C levels in claims data, right? So that's where the data comes from.

And I just want to make another point about the lack of risk adjustment for social factors, is that CMS is now actually paying for non-medical benefits, recognizing that if, you know, you live in a food desert it's going to be harder to control your HbA1C, you know, all those reasons. So they're paying for it because they recognize the impact on quality but not yet recognizing the impact on claims.

CO-CHAIR CELLA: Okay, so, keeping in mind the context, which is that these passed, and yet Jack has raised an issue, you're welcome to comment on that, particularly if you feel that

what Sherrie is suggesting, which is moving more towards some kind of unified composite, can be enabled or perhaps can include down the road some more consideration of social risk analysis.

MS. BARTON: Thank you very much, Dave. I'm Mary Barton, Vice President for Performance Measurement at NCQA, and I want to thank Sherrie for that thoughtful set of issues.

And I think, you know, there are a couple of composite measures. Of course, Minnesota has one, and then our comprehensive diabetes care, which I think the person on the phone was trying to suggest, is also a composite. It's not exactly a composite, but I think I would like to talk more in the future about how the individual error ties into how a composite could be a better reflection. And I'm curious as to how much this depends on measuring clinicians versus groups versus plans. But I would welcome the opportunity to talk more with you about that.

MEMBER SIMON: Excellent. By the way, I wasn't soliciting consultation from NCQA --

(Laughter.)

MEMBER SIMON: That's not me promoting business. But, so, the idea that, you know, that plan level issues -- the whole nest of environment of the whole healthcare system, it creates problems for us interested in measurement, because at each level there's error.

The question is, there are a couple ways to improve the precision that can actually make maximum use of multiple things measuring the same thing. And it actually helps you and reduces the standard error. When you get, you know, the adjustors in right, and all that other stuff, but it really does -- these are the kinds of things that, going forward, people with rich data sets like you have, that many people don't; claims databases don't have that kind of rich data; you can actually test and see at what point a measure becomes more characteristic of a provider or more characteristic of a plan and make adjustments in the kind of performance profile you create using different measures.

So it's just an opportunity, I think, to kind of help the whole field move forward in how we're thinking about the attribution of quality measures to a unit of analysis. Thank you.

CO-CHAIR CELLA: Well, I'll take you in one second. I apologize. I should have given you the chance to introduce yourselves to us. And I gather there's somebody else on the phone that we heard a little bit from. So just a quick introduction there.

MS. BARTON: Oh. Vice President of Performance Measures at NCQA, and I've been in that role for about eight years. And so I just wanted to talk a little bit about the other concern about case management --

CO-CHAIR CELLA: Before you do, there are others.

MS. BARTON: Oh, I'm sorry.

CO-CHAIR CELLA: You're with other people.

MR. REHM: Hi. I'm Bob Rehm. I'm the

Assistant Vice President for Performance Measurement. I've been with NCQA for nine years. I've been in NQF land for nine years, enjoyed every minute of it. And we have presented, Mary and Dan and I, at almost every standing committee that NQF has pulled together.

So, we really appreciate it. This is a great opportunity for us to get feedback and to help us, as we mentioned in the opening comments, to stay ahead of the puck as opposed to chasing it. Thank you.

MR. ROMAN: My name is Dan Roman. I'm a director with Performance Measurement. I've been with NCQA for just over seven years.

CO-CHAIR CELLA: And there was someone on the phone?

(Simultaneous speaking.)

CO-CHAIR CELLA: Okay. Go ahead, ZQ.

MEMBER LIN: I just had a technical question and I --

CO-CHAIR CELLA: Can you speak up a little, ZQ?

MEMBER LIN: I just had technical question about your measure, how you determine numerator comprised. So you based on two data, so one's administrative code, one is medical chart review, right?

For administrative code, it's okay, based on three categories. For medical chart review, you require a distinct numerical result. So I just wonder why you require a distinct numerical result from chart side, but it's okay from coding side.

MR. ROMAN: Sorry, is the question why -- so, the administrative codes we allow for this are CDT Category 2 codes that are specific to HbA1C results that either qualify or don't.

MEMBER LIN: Right.

MR. ROMAN: And then the results we look for in the record, it's the same set of requirements for what counts or doesn't. I'm sorry, maybe I didn't understand the question.

MEMBER LIN: So in a medical information form, say, when you're getting

information from medical chart you require, you know, a hospital provide these things in numerical results. So it kind of a different standard when you're getting information from medical chart compared to when you're getting information from an administrative claim.

MR. ROMAN: Okay, maybe this is helpful. Before PQRS or before PQRI and now PQD, which is the programs creating these Category 2 codes or G codes, quality measures, to make it simpler, primarily for clinicians who would be looking at their A1C of 6.9 or whatever. And then they would code. That would be coded as a G code or a CPT-2 code.

So you're right. It's capturing two degrees of separation from the actual lab value for the A1C and the chart, if they did a chart review.

The reason -- and it's just a few years ago that we accepted CPT codes. It has a certain -- there's a lack of -- they see this as check-the-box measures, as we agree in part. But

on the other hand, clinicians who are involved in the QPT program now -- and many people who record these measures -- they just say it's just so much easier if you'll just give us credits back.

So even though these measures are not provider level measures, they are at the health plan, we do accommodate that. We don't get a lot of this, as I understand it, necessarily from that, but it's to accommodate providers' burden, essentially. I hope that helps.

MEMBER LIN: Okay.

CO-CHAIR CELLA: Any further discussion? Patrick?

MR. ROMANO: Yes, hi, Patrick Romano. Obviously, NCQA has been the leader in this field of population health management measures for diabetes, so I have to give NCQA a lot of credit for really developing these measures and maintaining them over time.

But, you know, obviously, the bar keeps getting higher as we go through years and years of this process. And I think all of

understand that some patients with diabetes are just easier to manage than others. And, in some cases, it's fairly easy to get the A1C down under 9 percent, 8 percent. In other cases, it's extremely difficult, and we effectively can't do it within the range of what's tolerable to the patient in terms of medication side effects and medication burden and so forth.

And particularly in the Medicaid market, we see -- I don't mean to be critical of plans, but we see that these are very high risk, vulnerable patients. And there are incentives to skim the claim, so to speak, where certain plans cater to patients or enrollees in communities where they have access to more resources, where, frankly, it's easier to manage the diabetes.

So I would join with others in encouraging you to consider ways of incorporating risk adjustment into these sort of intermediate clinical outcome measures going forward. You know, if not this cycle, but the next cycle and beyond.

MEMBER SIMON: This is like pie-in-the-sky stuff, but as we're considering these kinds of opportunities, there are a broad -- we actually looked at nine measures in the 2009 article in Medical Care, nine measures of performance, including did you look in the eyes, did you check the feet, blah-blah-blah-blah-blah.

So, a suite of measures that you could then look at the attribution at the plan level, adjusted for whatever risk factors that are reasonable at given units of analysis, and you could actually help the field understand which are right for which purposes and which are not so good for other purposes, and do it empirically in a way that I don't think very many other folks can, because you have market, the medical records.

CO-CHAIR CELLA: Well, that was a great discussion. Thank you from our side as well for all the work you've done in this area.

So I think we don't have an action item here for this committee. The measure's

passed. But we are referring this to NQF and to the Standing Committee to work with you and encourage this kind of enhancing of the measures over time. Thank you for coming.

(Simultaneous speaking.)

CO-CHAIR CELLA: Then I'll turn it over to Dave N. for the afternoon.

CO-CHAIR NERENZ: All right, well, I'll try to continue this run of being on-time. Sherrie, thank you. This was a good way to tee this off, a good way to efficiently discuss it in the context of the three measure. Appreciate that.

Okay, so we're actually staying in a similar vein. We are moving now to Measure 0425, if you want to bring that up. This is functional status change for patients with lumbar impairments.

I think, again, I might be corrected here by staff, we're in a similar situation in that this has passed, but it was brought up by request of one of the members to discuss a

particular point about it. We didn't bundle it with the previous three because it's a different point.

So, I wonder, is it worth just a very brief staff flyover this one just so we get all calibrated? But I emphasize brief, because we're not revoting this one again, as far as I know. And then we'll turn to Sherrie as to discussion later.

MS. JOHNSON: Hang on. I had confused Bob and Mary and myself completely. Which one are we on? I'm trying to tell them whether they should or shouldn't go. Which one are we on?

MEMBER FABIAN: It's a FOTO measure.

CO-CHAIR NERENZ: This is a FOTO measure.

MS. JOHNSON: Okay. I'm so sorry, yes.

CO-CHAIR NERENZ: Sorry about that.

MR. STOLPE: No worries. Thanks, NCQA. And Daniel's in the hot seat now.

MEMBER DEUTSCHER: Do you want me to sit over there?

MR. STOLPE: I think you're probably in a good way now, Daniel. If you'd like to move, you're welcome to.

Okay, just some high level overviews per request of the Chair. This is a measure of functional status change for patients with lumbar impairments. But I'd like to read the measure description, because it is a little complicated.

So this is a self-report outcome measure of functional status for patients 14 years-plus with lumbar impairments. A change in functional status assessment using FOTO lumbar PROM is adjusted to patient characteristics known to be associated with functional status outcome -- that means risk-adjusted -- and used as a performance measure at the patient level, at the individual clinician, and at the clinic level to assess quality.

So, this is a PRO-PM. So, as such, the data isn't from an instrument and needs to undergo a little bit more rigorous testing than some of our other quality measures that are

brought to NQF for consideration.

This was evaluated as high for both reliability and validity. I'll stop shy of going through all of the data elements and score level testing that were performed. But suffice it to say there were many tests that the developer put forward, as well as some fairly robust analysis at both levels for both reliability and validity.

And let's go ahead and leave it at that and hand it back over to the lead discussant.

CO-CHAIR CELLA: Sherrie?

MEMBER KAPLAN: Ready?

CO-CHAIR CELLA: Yes.

MEMBER KAPLAN: Okay, so I have, first of all, in terms of reliability specs, Section S-6 states that the target population is age greater than or equal to 14; and then S-15 has the inclusion of patients under 8.

So, first, can you clarify the measure specifications? Which population are we talking about? Is 14 and over? Or do you actually

include both, under 8 --

MEMBER DEUTSCHER: Yes, I think we also responded to that in the response to --

MEMBER KAPLAN: Oh, and I didn't read it?

MEMBER DEUTSCHER: But it's a simple explanation.

MEMBER KAPLAN: Okay.

MEMBER DEUTSCHER: So I think what's in the specs relates to the general FOTO instructions when using a proxy. Actually, -- so it's not relevant for this measure, in the sense this measure is for patients 14 or above.

MEMBER KAPLAN: Okay.

MEMBER DEUTSCER: So we suggested in our response to add a note to that effect, to specify that this relates only to 14 and above, just to avoid this confusion.

MEMBER KAPLAN: And we came up on that faster than I was able to scroll down.

So the other question I had about reliability -- and this has already passed, so

this is not going forward again -- is that when you have multiple items per patient there's a within patient, across items error term.

And then you have, when you're folding it up across patients within a unit, you have another error term in the denominator. And then you've got the between -- the signal, between units.

So, going forward -- and a I don't think this is part of the guidance yet, so this is not fair -- but going forward, when you've got multiple item measures, that denominator error term has to include the patient across items when there are multiple items.

And it will come up in composites, no matter whether it's clinical or these kind -- doesn't matter what the source of the data is. Any time you create a composite you're -- if you've got a patient level composite this is a problem going forward in terms of precision and reliability. So, not for this analysis, but for going forward.

Then I had another question, and apologies if I missed it, Dan, because I --

MEMBER DEUTSCHER: Can I just comment on this before you go ahead? So, for a patient-reported outcome in general, the way the measurement works is, especially when you're using IRT and computerized adaptive testing. So we have an item bank and we do not administer all items to the patients. And there's a provisional estimate after each response. And so we accumulate those provisional estimates, and at the end we have all of these, and we can calculate a standard for measures which is the error associated.

One thing I wanted to point out, at least to my view, this is very similar to any measurement based on an instrument, whether it's a patient-reported outcome or not. So I'm not sure if the term composite is the one we want to use for a patient-reported outcome measure that has multiple items.

But anyways, as, you know, we

discussed just a few minutes ago, when we measure weight there's an associated error. When we measure blood pressure, there's an associated error. So I think your discussion point is relevant for any instrument that takes measure, not specific to a patient-reported outcomes with multiple items on it.

MEMBER KAPLAN: Doesn't have to be instrument-based the way the NQF uses instrument-based. It's any measure that uses a composite has associated error. And when you create a composite you just have to keep track of what that's doing to the standard error of measurement at the patient level, and then you fold it up at the next level, across patients within a unit and then between units.

And so what I think NQF is going to have to come to terms with is what constitutes the threshold of passing once you consider that level of error that's not really, right now, reflected in the ways that NQF has approved approaching reliability instruments.

But when the purpose of measurement is changing, and I know you're agnostic to the purpose of measurement, but when the NCQA's studying them, CMS starts adjusting compensation based on these measures, and the attribution to individuals units hasn't been well tested, I think that's a real caution that we need to express to folks who are using these for different purposes.

And that's where -- if you're just using these for quality improvement, you know, okay, we can all accept that there's error. But when you're starting to adjust compensation, that's when these estimates really matter because you've got to be fair about it. And that's where I think it's maybe time to start working with folks who are out there trying to do this the right way.

And I think he opportunities here may also exist for FOTO to make some, you know, testing --

(Simultaneous speaking.)

MEMBER DEUTSCHER: I guess the practical implication would be that, if we add an additional component to that denominator, we're probably just going to have to raise the bar in terms of number of patients assessed to make the measure reliable at the provider level.

I mean, from a practical perspective, that's going to be --

MEMBER KAPLAN: Well, it depends on whether each patient is giving you unique information. You know, if they're not, then that's not going to help you. But another way to do that is to add more measures.

You know, well, maybe the precision of the estimate you can improve by adding more measures. IRT doesn't allow you to do that, but there are kind of other approaches to measurement that do.

MEMBER DEUTSCHER: Right. And in this case, I'm not sure if this is an avenue --

MEMBER KAPLAN: Right. You know, you got what you got. And to Patrick's point -- can

I make one more comment? To Patrick's point, which I, you know --

CO-CHAIR CELLA: He stepped out.

MEMBER KAPLAN: So I'll make this in his absence. To Patrick's point, you know, we don't test the attribution, the amount of the variance explained. We actually did that in a 2009 article within the NCQA.

The attribution, at what point a measure in a value-driven analysis becomes more explained by patient level variation versus the unit, the doctor variation, for example. We did test that, and the hemoglobin A1C value came out 8, and the LDL came out 120. And so we fed that information back into NCQA and said, here's what it looks like when you actually test this.

We can empirically test this for most continuous level variables. Now, why wouldn't you do that kind of analysis, too? There are opportunities that folks have out there to begin to understand these thresholds and what they mean if you're going to use the measuring not for

taking care of patients, but actually for attribution for performance.

Those are different purposes when they're tested for patient care and empirical data guidelines and so on, but they're not tested on quality.

CO-CHAIR NERENZ: I just have clarifying and vocabulary question, mainly for Sherrie. It's about the word composite. And I want to make sure we're all on the same page, because I'm not sure I'm hearing it the same way.

My understanding of the word composite as used in this room is that there are two or more measures that have standing as measures. They may even be endorsed in their right. And they are combined, then, in some mathematical way to create a composite. So a composite is the combination of two or more measures.

Now, in this discussion, I'm not hearing that. The FOTO is a measure -- or this measure using that scale. There are a number of survey items, but they are not measures. So the

word composite, I'm trying to figure out how it bears on this discussion. I just want to be clear.

MEMBER DEUTSCHER: Well, this is what I mentioned before, that I'm not sure the term composite fits this discussion because --

CO-CHAIR NERENZ: No, I don't think it does, but I --

MEMBER DEUTSCHER: But I understand.

(Simultaneous speaking.)

MEMBER DEUTSCHER: I understand the concept that you're raising using the idea of a composite.

MEMBER KAPLAN: Yeah.

MEMBER DEUTSCHER: But, basically, relevant for this measure, it's maybe not the composite part of it but it's the error within the patient --

MEMBER KAPLAN: Yeah, but, so, this is really TMI for people who aren't fascinated by a measurement, but there's two arms of measurement models, a formative and a reflective model.

So, a formative model, say the stressful life mix. So, did you lose your mortgage, you know, did your spouse die? Did your kid go to jail? All of those things are individual components of something that's actually, if you add them all up, they shouldn't correlate with each other, you hope, at the patient level. But if you have one of those, it's bad. So that's like, when things cause stress. Those things cause stress.

But if you say, okay, my pulse is fast, I'm sweating, I'm this and that, you know, I have a racing heart and all of that other stuff, that's a reflection of stress.

The assumptions you make, measurement-wise, are different in how you approach those measurement models from a reliability and validity standpoint are different. But you're still adding them up. And so if you have a global concept called "I'm adding up a bunch of stuff," that's a composite. How you test the composite is different.

MEMBER DEUTSCHER: Well, but, again, I'm nervous about this, because adding up stuff can come in very different ways.

So, for example, if I have individual measures that have properties of reliability and validity and what not and I add them to make a composite, there are certain mathematical things that are happening and something of that involves the reliability and validity of the component measures.

Now if I take a bunch of survey items and created a score, and that score becomes a measure, I don't presume that each one of those individual survey items, itself, has reliability and validity. I generally don't test that. I may not even want to know that.

So it's different. It's not just adding up stuff. It's what you're adding. And I'm just -- all this is in service. I'm just trying to understand the point, which I believe is important, about the error variance, like which error variance and when does it matter and

when doesn't it matter. That's what I'm trying to set up.

MEMBER KAPLAN: Yeah, well, somebody like me, adding up stuff means something very specific. I think all of that stuff makes a construct. So I think all of that, those individual things, are related to some higher order construct, like stress.

You know, if you lose your mortgage, that's one thing. If your kid goes to jail, you know, you've broken your shoe lace for the 29th time, whatever all those things, those things, if you add up more of those, it's bad with respect to construct stress.

Like you're adding up stuff. But they have to have this underlying construct, otherwise there is no point in adding them up. You know, why would you do that?

On the other hand, and if you think about a math test, two trains left Chicago, you know, and then you've got a bunch of questions about some things, all of those things are

theoretically supposed to measure your math ability.

And, you know, I don't know if that helps us out of the conundrum of what is a composite, but for somebody like me, you'd better have a good rationale, theoretical rationale, for why those things should have come together. I'm sorry.

CO-CHAIR CELLA: This is -- maybe this is addressed mostly to the newbies as a welcome to the carousel. We've been around this carousel a few times and just jumped right on.

You know, I have had to induce some sort of self-inflicted electric shocks, in my terminology, as I've gotten acclimated to this committee.

So, for example, I think now, here, in terms of data elements and measures. And everything I used to call a measure is not a measure. It's a data element. Because I used to call patient-reported outcome measures, measures.

But I try not to do that here. They're

data elements. They're elements that go into a measure. And only measures, in this room, for me -- only measures, not data elements -- can be composites.

So, in this room, Sherrie.

(Laughter.)

CO-CHAIR CELLA: That's why I looked over there as I said that. I thought you might like it, but at least it's helped me. So if you're anything like me, maybe you were helped by that.

I think the point Daniel was trying to make is that the data element that is in the FOTO submission is more like hemoglobin A1C and blood pressure than it is like an SF-36 PF-10, because it's one number that keeps getting re-estimated each time you ask a question, maybe the same way that the laboratory configures out the hemoglobin A1C.

They've all got error. That was the point. All these data -- and you made that point -- all these data elements have error, but

there's not a multi-item in an IRT framework -- and Sherrie knows this -- in an IRT framework there's not a multi-item error, you know, Cronbach's alpha equivalent. It's one score.

I don't know if that's helpful or if that just spins the carousel more. But, for me, at least here, it's data elements are the things that go into the performance measures that sometimes are composites. I'll stop. Jack?

MEMBER NEEDLEMAN: Okay. I'm shocked that I'm the outlier here. But I am not incomplete on reliability. Insufficient on reliability.

I don't think we need to re-open the argument with respect to the endorsement of this measure. It's fine. But I do think this measure and the data that's been presented to us raised some long-term issues for the committee, both in terms of guidance -- in terms of our standards and in terms of guidance to developers.

So, one of the issues is that increasingly we're seeing measures that rely upon

the signal-to-noise ratio that John Adams put forward in an article, and we'll often cite that .7 standard in that Adams article.

But as John pointed out, in a simple binary classification, .7 produced 25 percent misclassification. So this committee needs to really figure out a way to say what level of signal-to-noise reliability actually we should expect to hit the standards that we expect for the uses of these measures?

And that's a long-term discussion. We're not going to get it done in the next two days. But we need to figure out a way to have that conversation and to move that forward. The white papers have been one of the vehicles for doing that. We should continue to do that.

But also we approved measures as specified, including the minimum number of cases. And what we're seeing in this measure, as we're seeing with others, is that the reliability is much lower, typically, for clinicians who have smaller numbers of cases.

And I think we need to also visit the reliability question with respect to minimum levels of reliability for clinicians or facilities with different numbers of cases in terms of assessing whether the cut-off for the number of cases is the right cut-off in terms of the specification.

And this measure raises those issues, I think, quite clearly. And that's one of the reasons why -- that's the principal reason why I said I'm not satisfied that I have sufficient information to make a judgment.

The other issue that this raises is about SES testing again. And this is something I would tell the committee to -- I would encourage the Standing Committee to reflect on.

We don't any direct measures -- we don't have a lot of measures in here of SES, but the acuity measure which shows to be very sensitive in the risk adjustment is time between the -- is actually what's called the QE is actually the time between onset of symptoms and

the initial evaluation.

And I can make up all kinds of stories about why that is SES-related. The developer doesn't discuss it. And perhaps the acuity measure as a risk adjuster sort of covers the sin of not analyzing more specifically the SES factors associated with delays in access to treatment. But somebody ought to think about delays in access to this treatment and how it's affecting our measurement of the effectiveness of the treatment.

The other thing, and this is a technical issue, was the treatment of the education variables in the SES test. And I read through the comments that were provided in response to that. They have left me now a little bit confused because what was said in the comments does not reflect what I saw discussed or the way the discussion was in the documentation.

So you've got these categorical variables for education. And, clearly, they're a bloc that's been broken up into categories. But

the assessment of do they come into the model or not is an assessment of each individual coefficient, each individual way.

I did not see any data on the testing of the educational variables as a bloc coming in or out of the SES measure. Now, the response is, that's what was done, but that's not what's reflected in the presentation.

And clearly, when you've got these kinds of categories, we ought to be seeing analysis of them as a bloc with expectations. So I also sort of comment that we don't see a trend to the data.

(Simultaneous speaking.)

MEMBER NEEDLEMAN: Well, it's not a perfect trend, but I do see a trend in the data, which suggests that those of higher education seem to have better resources than those of lower education. It's not perfect. The numbers are not moving consistently up across the levels as you move to higher level education. But I think if I did a regression line in that, I would not

get a slope of zero.

So, again, I think we ought to be giving our developers more guidance on what to say to us when they've got a categorical, and particularly categorical ordered variables, in terms of what analysis we'd like to see of those when they're talking about putting it into a risk adjustment model.

Those are my comments.

CO-CHAIR CELLA: Thank you. Again, we don't have a revote thing in front of us. These are general comments. If I could just take a second and make one quick observation and move on to the next measure also.

I really appreciate it, as I look at the discussion guide, looking at page 60, for those of you that have it up in PDF form. The column that shows variance explained at the provider level. To me, this is a nice capture of measurement property here of reliability. It also conveys to my non card-carrying statistician mind that we're talking about something here

that's about 6 percent signal and 94 percent noise.

Now, others may not view it the same way. And I'd also say that this is not unusual. The hospital readmission measures have essentially the same property. And others, too. But, as we continue our work and go forward and work with developers, I think we just want to develop a starter consensus about, you know, what should a number like this look like?

And I reflect on Mary's excellent comment at our most recent face-to-face meeting about how, you know, this signal that we've got here is sort of everything about the provider, quality and some other stuff. And unless the risk adjustment is perfect, these are essentially high end estimates. Mary pointed out that if the risk adjustment isn't perfect for something, much of what you think is signal, in fact, statistically is signal, is not really quality. It's just something you haven't adjusted for.

So none of this changes, I think, what

we should do about this. I'm happy to see it expressed that way. And I just think we, as a group, working in collaboration with developers to try to move this whole field forward, should try to sense of is it a good measure. Is it useful, is it informative, is it fair when it's 6 percent signal and 94 noise? We'll see. Sherrie?

MEMBER KAPLAN: I think one of the things that FOTO did well is they actually also look at residuals. And I think the unexplained variance here is key to kind of understanding much of the concerns that we've been raising.

I think this 94 percent business is it's going to have to come back to NQF and what we will tolerate for attribution at a certain level when it's useful for some things but not others. And that's what I was talking about earlier to NCQA, about, you know, maybe these measures aren't so good at the plan level. They're great at certain levels, but not at other levels. And will the field tolerate that? You

know, can we say, "Not so fast, CMS"?

CO-CHAIR CELLA: All right, Jack, and then, to stay on time, we should probably move on since we're not revoting on anything. Jack?

MEMBER NEEDLEMAN: So, I approach -- I often approach these measures from the perspective of information and usability for the users. You can have measures with 6 percent variation explained that nonetheless work. And so the other things I'm looking for, which are also captured by the reliability measures, but not perfectly, is how much variation is there between, once you've sorted people on the measure, how much variation is there between the lowest and the highs? How much room for improvement is there?

And the other is -- which also relates to the sort of inherent variability here -- how stable are the rankings and the orderings that are implied in this data? If they are so unstable that, by random chance or just by the bad luck of the draw this year, you wound up in

the bottom quintile versus and next year in the top quintile because it's the luck of the draw, we don't have a reliable measure.

So I think we need to start correlating some of these summary statistics with issues of stability of the measure, coupled with an understanding of how much variability there is from the top to the bottom or the top quartile to the bottom quartile, or something to give us a sense about how much -- what this variation explained is actually translating to in terms of the raw numbers.

CO-CHAIR NERENZ: All right. Thanks, everybody. And I guess now we're putting Subgroup 1 in the rearview mirror. Thanks to all of you for that discussion and background.

We move on, now, to a couple measures -- actually, there are more than two -- coming from Subgroup 2. So you folks can get caffeine and alert and reading to talk.

The first one we have -- in fact, we've got two before the scheduled afternoon

break, maybe a half-hour each. We'll see how we do it.

The first one is 0696. This is from the Society of Thoracic Surgeons. A candidate composite score. There's the word composite. And we will have Michael tee this one up for us.

MR. ABRAMS: Good. And I'm page 18 of the PDF for those of you who are following along on the discussion guide.

So, this is a maintenance measure. And, arguably, both a hybrid of some outcomes as well as a number of process measures. And, indeed, it is a composite. And I think a very clearly defined one, as it involved putting together the scores from 11 individual NQF-endorsed measures.

So let me quickly refer you to those 11 measures. It's broken up into four domains, the first one being the absence of operative mortality. Because of the use of mortality and morbidity in a measure, that's why I'm calling it a hybrid. But there's also some process measures

in there.

So, the absence of operative mortality is Domain 1. Absence of major morbidity across five different measures: reoperation for cardiac reasons, renal failure, deep sternal wound infection, prolonged ventilation/intubation, cerebrovascular accident, all negative morbidity outcomes that are each individual measures.

And then, in Domain 3, in the process area, use of internal mammary artery as part of the grafting procedure being one of the process measures that lies beneath this composite.

And then, finally, for additional measures that are process measures that involve different pharmaceuticals, all indicated for a typical CABG procedure: beta blockade therapy, antiplatelet therapy, discharge on beta blockade therapy, and then antilipid medications.

So all of these 11 are rolled up into what might be considered an all-or-none kind of approach. So, all 11 have to be hit in order for the numerator to be endorsed, and thus for this

measure to be considered accounted or included as a successful, positive outcome/process.

All the measures are based on registry data.

Some brief notes about the specifications. They did do a weighting procedure. So even though there are 11 measures that lie beneath, they give much more weight to mortality. Eighty-one percent of the composite measure that's reported is based on the mortality score; only 10 percent on the morbidity; 7 percent on this mammary graft and only 3 percent on medications. So they did some weighting there in order to create the composite, not a simple sum.

The way that they derive the data, from registry, as I said. Level of analysis is group practice. Exclusions or contraindications for this mammary artery graft use, perhaps because it's not the first time the individual had surgery, for example. And also contraindications for any of the pharmaceuticals

that were under the final domain that I described. So there are some exclusions, although they report that these exclusions accounted for only about a three percent reduction in the something like 143,000 cases that they used to do their testing.

The subcommittee gave it a moderate rating on reliability. So, passed there. It's a Bayesian model. Signal-to-noise ratio was reported at .68. In any case, not what we need to discuss today as the committee gave it at a solid moderate rating.

But with regard to validity, that was a "consensus not reached." And here are some points about that. And I'll try to signal some things for discussion and then hand it to Jeff to discuss that further.

But the measure gap actually was quite narrow, one could argue. The performance from 2014 data that they reported directly in one of their tables gave a performance rating of something like .97 percent with a very small

standard deviation running about a .0092, so there's some question about is there a performance gap related to this.

Another point, the way that validity was demonstrated was with a Pearson's correlation between, not the measure itself and an external validator, but the measure at two different time points using stability in the measure as a suggestion of its validity, that Pearson's and Spearman's reported for that type of analysis between two annual periods was .6, .63.

They talk about face validity, as well, being relevant to the creation of their measure. And then, finally, they do something that -- they use star ratings to cluster the performance of the different facilities into one of three categories.

And I put in the discussion guide a table. Actually it's on page 19 now. And if you look at that, or I'll describe it to you as well, most of the observations are falling into Group 2.

So, from their 2012 data, 76 percent of their facilities were performing at the two-star rating. When you get to 2013-2014, that number increases to 84 percent. So perhaps there's some question about discriminant validity and the way that they've set that up and compared it to their actual scores. There's not a lot of difference between the star ratings, I think, is the point there.

And then, finally, I can point out, with regard to validity, were two final points. Creating of the concept, the committee had no concern with regard that, so there's no reason to discuss that further. There was a lot of underlying measures, presumably, were persuasive to you all. And as I said previously, missing data is quite a small issue.

So the thrust of the discussion will probably be on this validity piece. So, with that, I'll had it to Jeff next.

CO-CHAIR NERENZ: And just as a process reminder, because it's consensus,

obviously, there will be a revote on this one.

So, Jeff, if you want to sort of focus to steer the subgroup and the rest of us on the key issues that are then going to be up for revote, that'd be great.

MEMBER GEPPERT: So I want to start with this one issue that is kind of a focus of my evaluation. And this may reflect some confusion on what was submitted or may reflect some confusion about NQF guidance around evaluations.

So, in many respects, the methods and the reporting of the results are really exemplary. You know, some best practices, really, in my regard. A lot of the methodology has been published in peer-reviewed literature and vetted through some great significant actual use.

My main concern was that the data that was submitted for evaluation was from the 2012 to 2014 time period. And so that sort of raised a question in my mind about whether -- you know, does this sort of meet the criteria that the data

being used in testing really sort of reflects the data that is intended for the specification? Which, presumably, would be now, or approaching 2020 through 2022.

So it's just the time period. Normally, we talk about that in terms of does it cover the same patients or does it cover the same providers, but also does it cover the same kind of time and space as what we're actually going to implement in practice?

And we've talked about at various times that too often reliability and validity are treated as sort of static properties of measures. But, really, they're not. They're the properties of measures that are context dependent. And one of those aspects of context was actual use. And, over time, through actual use, either reliability or validity could change over time. And so it's difficult to assess today what the reliability or validity of a measure is based on data from seven years ago after some significant use.

So that was sort of my main question,

my main concern. But then the developer, in the response, pointed out some guidance from NQF about this. And so that made me look at that. And there's quite a bit of discussion in the guidance document about shifts in emphasis for maintenance measures.

So it turns out that, for maintenance measures, in evaluation of the measure, there's a shift of emphasis in terms of the criteria that are considered. It says, for maintenance measures there's less emphasis in maintenance on evidence, reliability, and validity. There's no change in emphasis on specification and feasibility. But there's more emphasis on use and usability, you know, demonstrated improvement in a particular performance gap, is my way to sort of describe it.

So then if you actually read, you know, what the guidance says about sort of the shift of emphasis for reliability and validity testing, it says that the Scientific Methods Panel will provide the Standing Committee with

evaluations and rating of reliability and validity for new measures or for previously endorsed measures with updated testing.

So it seemed like the data that was submitted was probably from a prior -- from the initial sort of measure submission. That could be wrong, but that was sort of my assumption.

But if that's true, if the data that was submitted was from a prior measure submission, the prior endorsement process, and the developer is simply asserting that there hasn't been a change, then according to this guidance, you know, we shouldn't even be looking at this.

So that was my first question.

MS. JOHNSON: Okay, yeah.

MEMBER GEPPERT: Sort of topic of discussion.

MS. JOHNSON: So, a couple things. You are correct. Right now, we do not have any requirement that says that every so often the testing has to be updated. So, we don't have

that requirement right now. And that might be something that you guys want to suggest that we re-look at.

So, this particular measure, I'm not exactly sure why you are looking at it in terms of what's new. There's a couple things that could have happened.

One is this may have been one that originally they didn't give us the composite information that we needed. And this time they did, so now you have new stuff to look at that was not looked at before. That might be it, because I know that's happened a couple times.

The other thing is it could have been we also -- and I didn't list it in all the things of the complex measures -- but we also have a little out to where we say, "and, on occasion, NQF staff may want you guys to look at a measure." So it could have been that.

And I don't know, Michael, if you have any intel on which of those two, as to why it's coming to the Methods Panel. Do you know which

is it?

MR. ABRAMS: Yeah, I don't know. I think it's the latter, though, because it was a composite measure --

MS. JOHNSON: That the staff just tasked the Methods Panel to do?

MR. ABRAMS: Yeah, for that reason.

MS. JOHNSON: Okay. It could have been that.

So, back to your question, do we require, you know, they did, looks like 2014 is the most current data. We can't right now, with their requirements, say that that's not good enough.

Most of the time what we see -- not all measures, but a lot of measures that are currently in use often get recalibrated every year anyway. The CMS measures, for example, they recalibrate those risk models and those kind of things.

So, by definition, those are always going to come back because there's always new

data and new loadings, that kind of thing.

I am assuming that STS had not recalibrated these models since 2014. And we did ask that. Does that help? It may not be satisfactory, but --

MEMBER GEPPERT: Yeah, I just, from a plain reading of what the guidance says, if in fact these results have not changed since the prior submission, it seems like, according to the guidance, that would be enough. There's nothing in the guidance that says we'll be making an exception if staff think it's the right thing to do.

MS. JOHNSON: Yeah, it's a different spot. So, yeah, we can. And sometimes we might do that if we are feeling a little bit like we need the extra oomph of the SMP to help us look at a measure, even if it has gone through before.

But we'll find out why. I don't exactly know.

MS. WILBON: Yeah, I think, yeah, I think this might have been a staff.

MS. JOHNSON: Okay, so this, Ashlie is thinking that this is probably a staff pull, that we wanted you guys to look at it for us.

MEMBER GEPPERT: Despite that, one of the areas of emphasis has to do with performance gap and sort of where we pointed out that it seems problematic even in the earlier data. The 2014 data performance gap was not large if you sort of looked at, you know, extrapolated a trend, you might think that now there's probably no performance gap whatsoever. And so that, by itself, would be sort of enough to call it into question.

MS. JOHNSON: And let me interrupt. I'm sorry, Jeff. I know in the discussion guide the term performance gap, but I'm assuming that information came from the Meaningful Differences Section. Is that where that came from?

MEMBER GEPPERT: Yes.

MS. JOHNSON: Okay.

CO-CHAIR NERENZ: Okay, let's open now to the -- just to remind you, the subgroup

members here are Daniel, Jack, Larry, Joe Hyder, Alex, and I guess we have Eric on the phone.

So this is now our eligible discussants. And let's see who is up. Daniel was up first, then I got Larry second, and then Alex third.

MEMBER DEUTSCHER: So I guess my first point is also a question to NQF, and it regards empirical validity testing. So for such a composite measure that is up for re-endorsement, is that a new requirement that might not have been there when it was submitted previously? Is that something we do need to look into more deeply, or not?

MS. JOHNSON: This one, no. That would not have been new. No. So the only thing that's changed with validity is we have said that we used to take face validity kind of across the board and it didn't matter when. But now we say face validity for maintenance measures, we really don't like that very much. We'd really like to see empirical, but this is kind of the typical

way that STS does their validation. So I'm pretty sure that this is how they had done it last time, so this is not different.

MEMBER DEUTSCHER: So I guess my follow-up question is: is there any concern about the method used here for testing empirical validity, which is more of a longitudinal stability over measure than a validity against a different measure.

Now I looked at the response from the developers, and they give some explanations on why that happens, and if I understand correctly, their main point was that we kind of exhausted the other measures that we could look at because they're all within the composite. So I guess that might be a valid point in that case.

Is this longitudinal stability something that holds, or should they be looking at other ways of looking at validity, like known-groups validity, or other validity testing?

MS. JOHNSON: And I suspect that that is part of the reason that this was a stack pull,

that we wanted you guys to talk about that methodology, and weigh-in on whether you think that's a valid methodology for validity.

CO-CHAIR NERENZ: Okay, just a quick response. I'm going to put out for hypothetical --- you could probably have a totally invalid measure, meaning it didn't measure what you thought it did, but it measured something, and that would be stable over time.

MEMBER DEUTSCHER: Right.

(Simultaneous speaking.)

CO-CHAIR NERENZ: So okay, that's just my quick response there.

MEMBER GLANCE: So you know, I wanted to echo the comments that Jeff made about, you know, the STS measures are some of the best measures that we have that we've seen on this panel. They really do follow best practices. This is sort of an editorial comment. But a couple of the issues that were brought up, in terms of validity testing, first about the measurement gap, or lack thereof.

I think it's important to understand that at one extreme, if you have really, really poor risk adjustment, you're going to see a big gap. Okay? Because you're going to have differences between hospitals or providers in cases, so some hospitals are going to have really sick patients, some hospitals are going to have really healthy patients. And so you're going to see very different outcomes just because different hospitals have different severity of disease, and you're not risk adjusting for if you have a bad risk adjustment model.

Now, when your risk adjustment model is really good, okay, you're going to adjust for all that stuff, and all that's left is really differences in providers. And it may be that when you're comparing cardiac surgeons or hospitals that provide cardiac surgery, you have a pretty uniformly high level of care. And so there aren't really going to be huge differences, but there are still differences.

And as a potential patient, and also

as someone who -- CMS, you still want to be able to identify even relatively small differences between providers because that may help you make a decision about which cardiac surgeon or which hospital you go to, even if those differences aren't very big. So the fact that there isn't a really big measurement gap isn't enough to say, you know, this model, this measure is not valid. That's my first point.

The second point, in terms of using stability as a way of validating a measure, in some ways, measure stability may be one of the best ways that we have.

So the idea is that, you know, you look at your -- say your risk adjusted mortality rates, just to simplify, in 2016, and you look how well they agree with risk adjusted mortality rates in 2017.

If they don't agree very well, then you're sort of like thinking, well, you know, what information did those risk adjustment mortality rates in 2016 really give me if from

year to year they just fluctuate wildly?

If on the other hand, they show a fairly high level of agreement, that I think gives you more confidence in the validity of the outcome measure itself. So I would say that stability is actually a fairly important way of establishing validity.

And then the third point is about the fact that this was based on old data. And I agree with that criticism. I actually -- instead of giving this a high for validity, I gave it a moderate based on the fact that it was based on old data.

I actually really think that the measure developers have newer data. I'm not really sure why they didn't submit that information. And I think we should just ask the measure developer that question, you know? But I would certainly not fail this measure on validity. This is a very, very good measure, and I think that for all the reasons that I gave.

I think, one, the fact that again,

there is a measurement gap -- it's not a big measurement gap, but that's not unexpected. Two, stability is a good way to look at measure validity. And three, this is not old, old data, but it's not as recent as it could be.

CO-CHAIR NERENZ: Alex?

MEMBER SOX-HARRIS: Okay, thanks Larry. So what made me dissatisfied with the age of the data was that they reported some new performance results that were higher than the older data, and also more constrained. So if those things changed, then reliability as calculated has probably changed as well, and I just wanted to see it updated.

But I think more importantly is the lack of a performance gap speaking. A number of people have raised that issue. And has the measure really topped out? And so when they translate score to stars, I think there's an overstatement of difference, potentially.

So what they do, if I understood correctly, was if your confidence interval for

your performance crosses the average, you're a two star. And if your confidence interval is wholly above the average, you're a three star, and so forth.

I think the problem with this is you could get two facilities that -- or one facility that's higher, that has a greater confidence interval, that is rated a two star, versus a facility that's quite large and has higher volume, and that's lower rated, but it's tighter and does not cross the average line, and be a three star. And I think that's potentially misleading, and also I think it makes the spread, when you look at the star distribution, potentially bigger than it actually is.

So those were my two main concerns, other than the issue that's been brought up about there not being an empirical validity test that's due to some external reference. So I would respectfully disagree, but I think stability's important, but that's reliability. It's how well does something relate to itself? It's not a true

measure of validity. Those are my concerns.

CO-CHAIR NERENZ: I don't see any other of the name panels up. I'll just sort of make a last call for members of the subgroup, since we've got -- at least in the first pass, three people said low, two said high, one said incomplete. Do you want to ask any additional questions of each other to clarify or expand on anything? Because if not, then we'll open it up to the whole group.

(No audible response.)

CO-CHAIR NERENZ: Oh, I'm sorry. I'm reminded they have a developer representative on the line. This would be a chance to offer a response. Sorry, I didn't see you sitting here.

MR. SHAHIAN: Yeah, Dave Shahian and Mark Antman from STS are on the line, and --

(Simultaneous speaking.)

CO-CHAIR NERENZ: Okay, thank you. Just one second, we're having horrible volume problems. We can barely hear you, and I'm not sure we can control that, or if you have any

chance to speak up more at your end. We're working on it. Just hang on a second.

MR. SHAHIAN: Okay.

CO-CHAIR NERENZ: Try again.

MR. SHAHIAN: Okay.

(Simultaneous speaking.)

CO-CHAIR NERENZ: Better, thank you.

MR. SHAHIAN: Can you hear me now?

CO-CHAIR NERENZ: Perfect, thank you.

MR. SHAHIAN: Hi, it's Dave Shahian, chair of the Quality Measurement Task Force for STS, and I'd like to respond to some of the excellent questions that have been raised.

First of all, we did provide yesterday updated reliability, and what I would consider to be some internal validity testing, and I hope everybody has that. It's based on 2018 data. Did people receive that?

MS. JOHNSON: Dave, this is Karen from NQF. No, we weren't monitoring yesterday, so we have not made any of that available to the panel.

MR. SHAHIAN: Well, would it be

possible for you to look right now in your emails? Because this I think would settle a lot of these issues, if you would want to take a look, and then distribute that very quickly to the group?

MS. JOHNSON: Yeah.

MR. ANTMAN: And this is Mark Antman at STS. Just to clarify, your staff has just confirmed with us that a document that we sent actually very early this morning was distributed to the panel.

MS. JOHNSON: I think we're confused. We distributed something that you sent last week, but we have not distributed anything that came in yesterday.

CO-CHAIR NERENZ: How about if we do this? To our STS folks, if there are other points you can speak to, essentially just as a time holder, we'll see if somebody in here who has laptop access to the projector and the screen can connect all those things together.

MR. SHAHIAN: Okay. All right. Let

me talk while you try to get the data, but I will tell you a little bit about the 2018 data that we sent you. It's based on 152,446 records from 962 adult cardiac participants.

In 2018, in this particular heart rest period, the number of two star or as-expected programs with 80 percent one star, 4 percent, and 7 percent were three star. Over time, what has happened is that exactly what we would hope would happen in a very robust, data-rich environment, where we're providing feedback reports to programs, where we're publicly reporting -- you gradually get a convergence of scores.

Having said that, the 11 percent or so one or three star programs, it's still substantially higher, for example, than most CMS measures. It is still a sizable proportion, but as somebody commented earlier, you can't really survive in cardiac surgery unless you're a two or three star program.

Those programs either tend to improve, or they just die. And so to me, this number is

consistent with the number of high and low outliers in most other measurement programs that I'm aware of. And I think it also demonstrates the fact that we've been at this a long time and really been pushing for exactly what you're seeing here.

As a matter of fact, I have often commented that if we could have every program in the country statistically indistinguishable and functioning at a very high level, that would really be the ideal situation.

We're not looking to make somebody an outlier, we're looking to improve the quality of all programs in cardiac surgery so that patients then feel comfortable making their choice based on other considerations like geographic convenience and that sort of thing. So that, in terms of number of outliers.

In terms of the signal-to-noise ratio, signal-to-noise, reliability -- based on 2018 data, if you look at all participants, the reliability is 0.64. If you look at participants

with at least 50 eligible cases, which would be almost all programs, it's 0.66, and if you look at participants with at least 100 eligible cases, it's 0.67. If you look at most measures in surgery that I'm aware of -- and frankly, surgery does much better than medicine in this respect -- but even if you look at most surgical procedures, certainly most general surgical procedures, this is a very high reliability. And this is -- to my way of thinking, confirms the reliability of this measure.

And finally on the issue of the validity, it is correct that we have a bit of a problem in that we've used -- the major markers of priority are all already combined within our composite, so it's a little hard to get a good external gold standard to compare it to. We are, in fact, the gold standard, so one thing we thought to do is to see if in the construction of the composite, whether we have done anything squirrelly, would then make the domain scores seem unreasonable, or lacking in validity.

And if you look at -- and I don't know if you've received the information yet, but I'll just read it to you. If you look at mortality for one, two, and three star programs, for a one star program mortality is 4.3 percent, two star, it's 2.3 percent, and three star programs, it's 1.2 percent. Though it's really a threefold variation in mortality between one and three star programs. It's exactly the same for morbidity. It goes from 18.7 percent for one star programs, 11.2 percent for two stars, and 6.7 percent for three stars.

Even more striking for IMA use, one star, no IMA, 3.3 percent -- and for three star programs, it's 0.4 percent. So they're really an eightfold difference. And for use of all four NQF-endorsed medications -- one star programs, 19 percent did not receive at least one of those medications, and for three star programs, only 1.9 percent, so a tenfold difference.

So we think that in the construction of the composite, we certainly have retained a

kind of spread that we wanted to see in mortality, morbidity, IMA use, and medication use. There's a very substantial spread between one and three star programs.

Granted, there is some circularity in that argument because it is these measures that make up the composite, but I think, as all of you know, when you construct a composite, sometimes some untoward things happen, and you may not have the kind of relationship within the individual domains that you wanted. Here, I think we've demonstrated that we do.

CO-CHAIR NERENZ: Thank you.

MR. SHAHIAN: So I think I'll stop there.

CO-CHAIR NERENZ: No, that's good, thank you. That was very much on point. I'm just looking around the room, particularly at the subgroup members. Does anyone have any specific, additional questions for the developer?

(No audible response.)

CO-CHAIR NERENZ: I don't see any

hands up. Then let us then open it up to anyone else in the room wanting to discuss, make a point -- yeah, back to Larry.

MEMBER GLANCE: Just one more question for the measure developer. One of the issues that came up here in the discussion regards to the data that was used that -- how old the data was for the individual measures, in terms of their performance. So this is a composite measure that has mortality, that has morbidity. Did you look at the performance of those individual components?

So for the binary outcomes, did you look at the c-statistic and calibration in validation data sets using more recent data than the data that was used to create these models originally?

MR. SHAHIAN: Well, we have --- we've redone all of our risk models, and the performance of those models has all been very high. I'm not sure what else I can say. We work constantly doing that, and we just published the

2018 update of all our risk models. And in every case, except for I think internal infection, the c-statistic was improved over the previous models upon which we originally developed this composite.

And somebody had asked about re-calibrating the models, too, and I just want to point out that our statistical team does re-calibrate our models every harvest period, so that's at least twice a year. And they're re-calibrated to get an ODE ratio of 1. Usually that re-calibration factor's very small. It's, you know, 1.00-something. But they are re-calibrated with each harvest.

CO-CHAIR NERENZ: Thank you. And so --

(Simultaneous speaking.)

MR. SHAHIAN: And the other thing I forgot to mention was that our statistician has looked at the correlation between composite score estimates over two successive time periods, and that correlation is 0.63, if that helps.

CO-CHAIR NERENZ: Yeah, it helps.

MR. ANTMAN: This is Mark at the STS. May I just add to what Dr. Shahian said that the updated adult cardiac risk models that were published recently, we did send those papers from the STS annals of thoracic surgery, along with our templated response tied to the Methods Panel staff, a couple of weeks ago. So the staff and hopefully the panel should have those updated risk models available to you.

CO-CHAIR NERENZ: Got it, yes. Yes, thank you. All right, I don't see hands up or raised eyebrows or name tags up, so I think we may be ready then to call the question here.

So if you're on the subgroup -- oh, Sherrie?

MEMBER KAPLAN: Just one point of contention with Larry, that we've been back and forth about, but stability is not a measure of validity because my bathroom scale tells me the same thing every moment. It is wildly wrong, but it's completely stable. So just to note that --

MEMBER GLANCE: How do you know it's wildly wrong?

MEMBER KAPLAN: It's wildly wrong because I go to the doctor, and he tells me how much I weigh.

CO-CHAIR NERENZ: All right, well, let's keep that in mind as we go ahead. So would members of the subgroup -- and you know who you are -- go to one link and one polling place? The rest of us go to SurveyMonkey, and let us do this then.

MS. JOHNSON: And let me remind you we are only voting on validity for this measure. This is the one where CNR reliability had passed, so we don't need to re-vote that one.

MS. OGUNBEMI: Okay. We are voting on Measure 0696, the STS CABG composite score. We are voting on validity. The voting is active for subgroup members. Your options are high, moderate, low, and insufficient.

So our votes are 33 percent high, 67 percent moderate, 0 percent low, and 0 percent

insufficient. And the counts are two votes high and four votes moderate. Measure 0696 passes on validity.

CO-CHAIR NERENZ: Thank you, everybody. All right. Pause to take a breath. Move on before our break.

We are now moving to -- same subgroup, Measure 3537. This is a measure of intraoperative hypotension. It comes to us from Mathematica. Staff name is on my agenda.

(Laughter.)

MR. LYZENGA: So yes, the 3537. This is a measure of the percentage of non-cardiac, non-emergency surgery cases involving general anesthesia, or monitoring anesthesia care, in which mean arterial pressure fell below 65 milligrams -- milliliters -- I actually don't know what that measure is -- for a cumulative total of 15 minutes or more. So this is sort of an intermediate clinical outcome, looking at a particular measurement within the surgery, making sure it -- of hypotension.

This is a registry-based measure intended to apply at the individual clinician level. They did a signal-to-noise analysis for reliability of the measure score using the Adams method. They provided the distribution of those reliability statistics by clinician denominators, sample size, and by quartile -- had, you know, some fairly high scores. One of the concerns that the panel members raised was they wanted to see that those reliability statistics were reported by case volume, and the developers did provide that in their follow-up material. We can discuss that in a moment.

For validity, they did a couple of types of validity testing. One was predictive validity, by which they examined the association between unadjusted measure score in several negative outcomes, and then linked to intraoperative hypotension.

Their results were as they expected. Patients who experienced those negative outcomes had significantly higher rates of hypotension

than patients that did not experience these outcomes. And then their second way of assessing validity was in known-group validity assessment, in which they tested whether anesthesia cases involving patient subpopulations known to be at greater risk for intraoperative hypotension had significantly poorer measure scores than those cases not involved with those high-risk subpopulations.

Their results were consistent with what they expected, with the exception of patients who were 65 or older, who had slightly but significantly lower intraoperative hypotension rates than patients who were under 65, which was contrary to what they were expecting. And with that, I will turn it over to our lead discussant, who I think is Jeff Geppert --- oh, Larry.

MEMBER GLANCE: Okay, thanks. Thanks for that really good summary. So this is a really interesting measure, and basically the idea is that if you have a low blood pressure

during an operation, you're more likely to have a bad outcome, you're more likely to die, you're more likely to have a heart attack, kidney failure, et cetera, et cetera. And to surgeons and to anesthesiologists who practice out there, I mean, this has a lot of face validity. You know, you don't want to go out there and be really hypotensive for a long period of time.

In terms of the empiric evidence supporting that, so far there's not a huge amount. Even though it really, really makes sense to most of us, the reality is that there is very limited evidence to support this. Most of this from observational studies, which really do not establish causality. There is one randomized control trial that was published in JAMA, but that looked at sort of standard blood pressure control versus -- which they defined as keeping the blood pressure within 40 percent of what your normal is, versus keeping it within 10 percent.

So very few of us would let the blood pressure drop to below 40 percent of what it

normally is, so that really wasn't a fair trial. This is sort of by way of introduction.

So this is a risk adjusted outcome measure. It's a very parsimonious model, and there's nothing necessarily wrong with a parsimonious model. But in this case, one of the -- basically it has age, sex, ASA physical status, how long the surgery lasts, and body mass index. Okay?

So one of the most important predictive risk factors is ASA physical status. And for those of you who are not anesthesiologists or surgeons, which I take it is the vast majority of us -- of you, not me. So ASA physical status, one, you're a normal, healthy patient. Two, you have mild systemic disease, like you're a smoker, you have some well-controlled hypotension. Three, you have severe systemic disease, so you have poorly controlled diabetes, or you're end-stage renal disease. And then four, you have severe systemic disease that is a constant threat to life.

So why am I going over this? Well because the ASA physical status is not as objective as it sounds. Basically, I as an anesthesiologist code that at the very beginning of the case. So if I know that my performance is going to be measured, that maybe, you know, people who are a lot sicker are expected to have a higher incidence of low blood pressure. And maybe I don't do such a good job every day, maybe all my patients have a low blood pressure, but I don't want to be singled out as being a low performance outlier. So I'm going to upcode all of my patients.

So this particular measure in terms of the risk adjustment is really, really sensitive to upcoding because the people whose performance is being measured are the people who are making the determination of the ASA physical status, which is one of the main risk factors in this model. So that's a big problem with this model.

The other thing is that it doesn't really account for the complexity of the surgery.

So if you're doing a mastectomy, you know, that's a relatively superficial procedure. People aren't likely to get hypotensive. If you're doing an open abdominal aortic aneurysm repair surgery, those are a lot sicker patients, there's a lot more volume shifts, much more likely to get hypotensive. That is not at all captured in this model. Okay?

So the model itself is a little crude, and to my way of thinking, it's not quite yet ready to be released in the wild. Okay? The second thing is that not only is the model probably not quite as good as it needs to be when you look under the hood, but it was developed using just two different centers, two different hospitals. Okay?

So if you look at the typical risk adjustment measures that we look at, the good ones -- you know, the STS measures 1,000 different hospitals, you know, CMS outcome measures, you know, 2,000 hospitals -- this one is based on two hospitals. Okay? So it is very

unlikely that you can generalize anything about this measure and about its performance, and its validity, and its reliability to anything outside of the very limited data set that was used to develop the measure.

And those were really the main issues that I had with this. I think that the reliability was quite good; it's just there are major problems with validity, and mostly around a, the risk adjustment model, b, the generalizability, and c, the very real potential for gaming.

CO-CHAIR NERENZ: Okay, thanks. We're going to do a couple comments, and then if we do have developers available on the line, ask them to respond. Let's do either building on those comments, challenge those comments. Alex, and then Joe.

MEMBER SOX-HARRIS: To build on the comments related to the contents of the risk model, there's another variable that I had a question about, which was the length of the

operation. Things that go in risk models should be before and related to the outcome, and not gameable. It's a clinical question, I'm not a clinician, whether length of the operation could be related to the outcome in a way that was, you know, in both directions.

So the longer the operation, the more opportunities would be for having low blood pressure, which might make the operation longer.

MEMBER GLANCE: So in a sense, you're right in almost all contexts. You do not want to put things that happen during the episode, like blood transfusion and stuff like that, into the model. On the other hand, in this particular case, if you're doing a really long operation, it's more complex possibly, and you are maybe more likely to have episodes of hypotension.

So if you were comparing someone who routinely does really long, complex procedures to someone who does 30 to 60 minute procedures, you would want to take into account for the complexity and the duration of the surgery. I

don't think that's completely unreasonable.

MEMBER SOX-HARRIS: Yeah, so perhaps a way around this concern is to have the average length of the procedure -- not the actual length of the specific procedure -- included in the model, which would account for complexity and length, but would not have this issue I'm worried about.

MEMBER GLANCE: Agreed.

CO-CHAIR NERENZ: Joe?

MEMBER HYDER: Larry made some excellent points. I'll make a couple others. This has excellent face validity, but the challenge is really in the details, so one of the issues is that the measure as specified is at a level of clinician, not facility or otherwise.

For most anesthetics given in the United States, they're provided under a care team model, so there may be one or multiple anesthetists, one or multiple anesthesiologists. The developers have come back and said we would attribute the performance, and encased all of

those individuals. I think we'd lose some folks on the validity with that.

So even if one person controlled the blood pressure, they would get dinged for others, in the same case not controlling blood pressure. And then Larry made an excellent point about risk adjustment. So as far as the length of surgery, that's something that's determined after the case, so you could come up with a philosophical piece about: should we use something that occurred after the case to risk adjust?

Most of us would say no. I think the exception would be looking at what NISQUIP does, and also what STS does, which is they use a surgical CPT to risk adjust. So the CPT may be pre-specified, but the one that's documented in the risk model is the one that was performed. I think there'd be an opportunity to use something other than length of surgery with surgical CPT to try to get at a much more meaningful risk adjustment.

I think what happened in their

reliability calculations here -- and I could be completely wrong -- is that inadequate risk adjustment put variability in the hands of individual clinicians rather than back into the case. And clinicians take care of certain kinds of patients for certain kinds of cases in a non-random way, so you're going to have a falsely elevated reliability for those clinicians.

And then another minor point is I think there's some exclusions here. They talk about monitored anesthesia care, general anesthesia, and so forth. Some of our surgeries require induced hypotension, and it's not a rare thing. This includes, for some surgeons, all their joint replacements, a very common surgery, spine procedures, also very common surgeries. There's no exclusion or a way to get around induced hypotension at the request of the surgeon.

So I think there's some important limitations that may have affected the reliability and the validity for this one.

CO-CHAIR NERENZ: Thank you. I'm just going to make a quick observation just to kind of guide us as we keep going here, and then we'll go to Daniel, and then I'm going to go to Mathematica.

Jeff points out -- and I think it's good for us to keep in mind -- there's this really narrow hair split distinction in some of these areas, say of risk adjustment, where whether it's our call, or is it the Standing Committee's call? And I just want us to be thinking of that -- that although it's sometimes hard to know exactly, we've often said let's focus on the math, let's focus on the statistics, let's focus on model construction, that kind of thing.

If a certain variable should or should not be in because of its clinical meaningfulness, or is it specified a certain way, we typically -- at least in our young history so far, have said that's maybe something up to the Standing Committee. So I don't think anything to this

point has been inappropriate, but as we go from here to the voting point, let's make sure that we don't put ourselves in the position of the Standing Committee to say, you know, conceptually or clinically or medically the right things are in there.

And I realize it's a hard thing to keep, but that's -- okay, so Larry, you want to speak directly to this point? And then I'm going to go to Daniel.

MEMBER GLANCE: So you know, I think this is a bit of a philosophical point. And I think sometimes it's very difficult to separate out methodological expertise from content expertise. So I think in this particular case, certainly some of us have very, very specific content expertise.

But, even in a more general sense, when you're looking at the validity of the measure itself, I think part of that involves looking under the hood and looking at the risk factors that were included, whether they were the

right risk factors or not the right risk factors. So we talked a little bit about this. You know, if you're going to look at an episode, are you going to include complications or not? Should you be including hospital characteristics, or not?

So you know, in this case, should you be including a variable that we know is very easy to game? And we're telling you that right now because we have that content expertise. So I actually think that we ought to be careful about not making a strict kind of black and white, sort of like, divide. You know, I think we ought to be looking at the risk factors. In some cases, it's very appropriate to do so.

CO-CHAIR NERENZ: And the process allows the Standing Committee to overrule us anyway. So okay, Daniel, and then Jack.

MEMBER DEUTSCHER: And just a quick note about reliability. They did provide some additional information that we requested. And so I'm looking at this additional table that splits

the clinicians by sample sizes. And if I'm reading this correctly, although the overall reliability was very high, I think probably one of the reasons was that most clinicians have very large volumes.

But there was still one group of clinicians with one to 30 cases that had an average reliability of 0.5. So and it wasn't a very small group. It was about 12 percent of the sample. So I think maybe for reliability it should be noted that the measure isn't reliable, but above this threshold.

CO-CHAIR NERENZ: Yes, thank you. Jack, and then back to Joe.

MEMBER NEEDLEMAN: Daniel, where was that report of the reliability for cases one to 30?

MEMBER DEUTSCHER: It's in the PDF in the discussion guide.

(Simultaneous speaking.)

MEMBER DEUTSCHER: In the discussion guide.

MEMBER NEEDLEMAN: Because that's not in the original evidence package.

MEMBER DEUTSCHER: Right.

MEMBER NEEDLEMAN: My comment about it was not in the original evidence package. They only report reliabilities for the whole sample going down -- and again, that represents an inability to provide data that's useful for assessing whether the cut point for minimum number of cases. And that made the evidence package insufficient for making appropriate decisions. And this table just reinforces and confirms that.

MEMBER DEUTSCHER: Yes.

MEMBER NEEDLEMAN: The conclusion should draw from the table on page 22. It's very different from what you would draw from the evidence package.

CO-CHAIR NERENZ: True. Yes, thank you. Joe, and then back to Larry.

MEMBER HYDER: I want to speak to your point because my interpretation is that it may be

deceiving. So if Larry and I both take care of very sick people, and he takes care of them for a really big surgery, and I take care of them for a really small surgery, and we don't risk adjust for that, what will happen is that it will appear with higher reliability that he's a lot worse than I am.

And so I think that the adequacy of the risk adjustment or the inadequacy of the risk adjustment in this case may be contributing to a falsely elevated reliability estimate. I could be wrong, but that's my distinct impression.

CO-CHAIR NERENZ: All right, Jeff?

MEMBER GLANCE: Just one quick note. This is actually something that we talk about in our white paper. The fact that before you look at reliability, you should be looking at risk adjustment because if the risk adjustment is inadequate, then you're going to have falsely elevated reliability. And I think that Joe spoke to that very well.

CO-CHAIR NERENZ: All right, Jeff?

And then I do want to let Mathematica get a chance here.

MEMBER GEPPERT: I'll be quick. I don't know, for some reason not feeling very positive today.

(Laughter.)

MEMBER GEPPERT: The way that, you know, the reliability and validity results are reported even -- you know, we asked for some additional results. Those were provided. Very illuminating. I really liked the way they did the validity. And I'm looking at some -- you know, finding the implicit quality construct with our measure, but then I'm looking at some downstream outcomes. You know, we don't see that very often. This is one of the few cases where the developer actually looks at, you know, the relationship with downstream outcomes, so I think, you know, in a lot of respects, they've done an exemplary job.

CO-CHAIR NERENZ: All right, on that happy note, Cindy from Mathematica -- oh no,

Patrick, but you're not in the subgroup. Are you directly on one of these points?

MEMBER ROMANO: I'm not in the subgroup, so I'll --

(Simultaneous speaking.)

CO-CHAIR NERENZ: Okay, let's do Mathematica first, and then we'll see where we go.

MEMBER ROMANO: Okay.

CO-CHAIR NERENZ: Cindy?

MS. CULLEN: I'm going to hand it over to our principal investigator, Dr. Anna Christensen.

MS. CHRISTENSEN: Hi, yes. This is Anna Christensen. Thanks for giving us a chance to comment. Let me start by talking about some of the statistical results that we provided, and focus more first on some of the more traditional things that the Scientific Methods Panel reviewed.

So as the discussants have mentioned, we did provide new reliability. People showing

reliability results above 0.8, and in many cases, above 0.9 for the vast majority of clinicians in our sample. As mentioned, we only have two hospitals in our testing data so far. That's because this is a measure that is not yet in use. It's under consideration for use at this point.

And we use data from the anesthesia information management system, so we don't have larger amounts of data at this point to test the measure. But in the sample that we did look at, it's high reliability for the vast majority of clinicians.

And then the last commenter mentioned the new validity testing data. I hope everyone had a chance to see that. We provide a graph that shows that the clinician level risk adjusted measure is very strongly associated with downstream adverse outcomes for patients. Clinicians with the worst two quintiles of this measure have higher rates of acute kidney injury, cardiac surgery, and death in hospital.

In terms of the risk adjustment -- the

items in the risk adjustment measure, the discussants mention the ASA status, which yes, is something that is assigned by either a nurse clinical anesthetist or the anesthesiologist. We were told on various site visits that we went on that that was something that, you know, although there is some subjectivity in assigning it, it's something that clinicians feel like they're mostly on the same page on. There's some research showing some EquIS inter-rater (phonetic) reliability on ASA status.

In terms of the surgical length, and that being something that is determined after the case, the length of the case is something that is used for billing purposes. So it's very reliably reported, and trying to mess with that length of a case between anesthesia start time and end time would amount to billing fraud. So we were reassured that that would be something that wouldn't necessarily change, or couldn't be gamed.

In terms of the induced hypotension,

that's something, right, the measure does not deal with, but some of our clinician consultants had mentioned that that is not something that is very common. Let me check with my team and see if there are other points that they wanted to address. Okay. Hearing none, if there's anything else that I have forgotten to address, please let me know. We'd be happy to talk through some more of our position.

CO-CHAIR NERENZ: All right, thanks Cindy. Our process here is now we open up to the entire panel for comments or questions. Patrick had the first crack at this.

MEMBER ROMANO: Yeah, I mean, two things. I mean, one unfortunately is that I certainly understand that the developers are trying to use the duration of the surgery as a proxy for the complexity of the surgery. Unfortunately, it sort of violates the rule, that it's endogenous, you know, because clearly a period of prolonged hypotension is going to affect the total length of the operation.

Either it'll lead to a termination of the operation, it may shorten in some cases, or it may lengthen in other cases because of the maneuvers that are being made to address that hypotension. So clearly, that is not an acceptable theory. Now, it seems to me that the developers have access to measuring the average duration of operations of that type, which would be exogenous. So if you took it away from the length of that particular operation in that patient, and used instead the average length of operations of that specific type, then you could resolve that problem.

The other thing I wanted to point out is that they cite Kendale's paper for predictive analytics, and the most important risk factor that was identified in that paper was actually the baseline blood pressure level. So I know there was some discussion in the documents about why that could or couldn't be included, but that would be an important thing to consider.

CO-CHAIR CELLA: All right, thanks.

Any other comments by other members of our group? Jack, you look like you're thinking about it.

MEMBER NEEDLEMAN: Yeah. I'm looking in the documentation. I'm not finding it immediately, so I'll ask the developer. In light of your reliability findings, is there any minimum number of cases that are specified for inclusion in this measure?

MS. CHRISTENSEN: Right now, we don't have a minimum case volume. The testing is all done with clinicians of all numbers of cases.

MEMBER NEEDLEMAN: Okay. Yeah, that might be a decision you want to reconsider, given the reliability numbers that you've presented to the committee.

CO-CHAIR NERENZ: Karen, let me just ask you, the general measure endorsement process --- clearly this matters. We all see it in the table. There's no mystery here. If there should be a minimum, is it up to the developer to state the minimum at the front-end of this? Can we, in our process, state a minimum? How does this work

now? Because the table says what the table says. It's not so hard to figure out.

MS. JOHNSON: You could suggest that they might want to consider a minimum, but they would be the ones who would actually --

CO-CHAIR NERENZ: So that's not up to us to declare?

MS. JOHNSON: Correct.

CO-CHAIR NERENZ: Okay.

MS. CHRISTENSEN: Yeah, we would be happy to make that change. It was our understanding that that would be made at the point of the program at which the measure was included, but we could also make that suggestion in our specification.

CO-CHAIR NERENZ: Yeah, I think part of the concern we've had, at least looking back historically, is that a measure gets NQF-endorsed under certain conditions, so then the users and programs go off and do whatever they want to with it. Different setting, different minimum sample, they do whatever they want. So the feeling here

is things have to be a little tighter than that. But that's okay. Christie?

MEMBER TEIGLAND: Just quickly. I don't know if you've looked at, you know, any ranges below 30, because it might not be 30. I don't know how you got, you know, those cutoffs, but it could be 20, it could be, you know, 10 maybe. But you might want to just do a little further investigation of that bottom number.

CO-CHAIR NERENZ: Okay. All right. I'm sensing a movement to call the question, and then break. That's good. Larry?

MEMBER GLANCE: I just wanted to make just a general comment to the measure developer. I actually think that this measure -- although in my judgment, I don't think it's quite ready to be released -- I really like the concept because it is very likely that there is an association, a causal relationship between hypotension intraoperatively and mortality, and major morbidity.

But I think that this measure needs a

little bit more work. I think you need to develop your models a little bit better, and you need to have more hospitals in your data set. I think that you need to consider other risk factors besides the ones that you've included, but I don't know how this group is going to end up voting. But if it ends up voting not to endorse this measure, I don't think in any way should that be seen as our group saying this is not something that has some promise.

I think this is potentially a very important measure. There are not a lot of measures out there that really look at what perioperative physicians do in the operating, what anesthesiologists do. So this is a really interesting measure, and I think it needs to be pursued.

MS. CHRISTENSEN: Thank you.

CO-CHAIR NERENZ: All right, let us proceed then. Subgroup members to your voting domain, and the rest of us SurveyMonkey, and let's do the business.

MS. JOHNSON: And we'll be voting on both reliability and validity on this measure, which --

MS. OGUNBEMI: So we are voting on 3537, reliability. The options are high, moderate, low, and insufficient. Voting is open.

Okay, voting results --- oh, it's moving. I'm going to close the responses now, thank you. The measure still passes, but the results for Measure 3537, reliability, 33 percent high, 50 percent moderate, 17 percent low, and 0 percent insufficient. And the count is two votes high, three votes moderate, one vote low, and zero votes insufficient. Measure 3537 passes on reliability.

And now we move to validity. The voting is open for validity of Measure 3537. Voting is closed. Let's see. So we have 0 percent high, 33 percent moderate, 67 percent low, and 0 percent insufficient. The counts are zero votes high, two votes moderate, four votes low, and zero votes insufficient. Measure 3537

fails on validity.

CO-CHAIR NERENZ: All right. Well thanks everyone for the discussion. Good work. We are now at a scheduled break time. The agenda says we're back at 3:00. That is a little challenging. We're close, though. How about if we try for five after, and I'll try to get us going as close as we can to that.

(Whereupon, the above-entitled matter went off the record at 2:53 p.m. and resumed at 3:09 p.m.)

CO-CHAIR NERENZ: Our speakers need to go green. All right, getting ready to go here in just a second.

Just to get us oriented, we have three measures to talk about in a little over an hour and a half, so I think we can do that. We are still under the auspices now of Subgroup 2, although we'll switch eventually to 3. We are now talking about Measure 0018. This is also an NCQA measure about controlling high blood pressure. This is actually in front of us because on

validity, it was a consensus not reached. So, Ashlie will tee this one up for us.

MS. WILBON: Yes. Thank you, Dave. As we've mentioned, we're looking at Measure 0018, Controlling High Blood Pressure, from NCQA. This is a maintenance measure. The description reads, the percentage of adults 18 to 85 years of age who had a diagnosis of hypertension and whose blood pressure was adequately controlled less than 140/90 during the measurement year.

We have categorized this measure as an intermediate outcome. It is a health plan level measure, as you are aware. It does not contain risk adjustments, and there is no analysis of social factors. And the developers described using plan type as a bit of a proxy for income and socioeconomic status.

As we mentioned, reliability passed with a rating of high. They did do the testing for reliability using the beta-binomial approach, signal to noise analysis. Overall reliability ranged from .98 to .99 across the three types of

plans.

Again, while it passed, there were a few concerns expressed by reviewers about the clarity and consistency of the specifications and various age ranges used throughout the specifications that seem to be inconsistent, and some clarity around how -- which target was used for blood pressure. But I think maybe we could get some clarity from the developers a little bit later on that. But overall it passed on reliability so we won't spend time there.

For reliability, work consensus not reached -- I'm sorry, for validity, work consensus not reached, so we'll focus our discussion here, I'll just do a brief summary of some of their testing approach and then means, and some of the concerns, then we'll turn it over to John for a brief summary as well.

So the validity of the health plan was demonstrated using -- with construct validity, the entire heated beta sample. The validity was conducted in correlation with another measure,

which was the comprehensive diabetes care blood pressure control measure, which looked at the percentage of adults 18 to 75 of age with diabetes, whose most recent blood pressure level was taken during the measurement year, was less than 140/90.

The Pearson correlation, they did a Pearson correlation between these two measures and across the three types of health plans. Those scored ranged from .75 to .93. Medicare was the lowest and commercial plans with the highest correlation score.

The concerns around validity are related to the lack of analysis around the multiple data sources, the exclusion of lack of a risk adjustment, and the comparative measure that was used to demonstrate construct validity. So, I'll stop there. John?

CO-CHAIR NERENZ: Daniel?

MS. WILBON: Actually, it's John, but if Daniel, I don't know if --- oh, I'm sorry, I'm looking at the wrong measure. Daniel. Sorry. My

agenda was on the wrong side. You're the lead discussant?

MEMBER DEUTSCHER: I'm not supposed to discuss this ---

MS. WILBON: Oh, okay. I can keep going.

MEMBER DEUTSCHER: Yeah, go ahead. I'll say what I know.

MS. WILBON: I'll start with the concept validity. There were concerns over the method that was selected to test the correlation demonstrate construct validity.

The concerns were around the fact that they both have, potentially have the same data elements around blood pressure measurements, as well as the overall measure, so that the diabetes population eventually would be a part of the overall population as it's measured in the packet of the measures that we're looking at. So there was concern that the measures aren't actually independent measures themselves, and that there may be some completing there with comparing those

measures and the direction of the quality performance score.

The next concern around risk adjustments, that there was not any risk adjustment, there was concern about this and whether the stratification alone of the health plan types was enough to account for the differences in the population that is generally factored with the different health plan types.

Additionally, there was some notes from reviewers around whether or not, in addition to socioeconomic differences, that there may be a need to adjust for additional clinical factors that may come into play with certain populations that may be more challenging than others to control blood pressure.

Multiple data sources, again, the measure as specified identifies at least three sources for collecting data for the measure claims, electronic health data, electronic health records and paper medical records. There was not any analysis in the submission, the initial

submission around how there may be differences in reliability or validity based on where the data was extracted from for the measure.

And then finally, some concerns were kind of questions from the reviewers about additional analysis for the exclusions to determine kind of a sensitivity analysis to determine that the percent of cases that were included actually aligns with what you might see in the literature or that they don't overall impact the performance scores in some other way.

So with that, I will pause and open it up to the subgroup.

CO-CHAIR NERENZ: Yes, why don't I invite anyone from the subgroup who wants to make a comment or questions to the measure developers. Okay, Daniel, yeah.

MEMBER DEUTSCHER: Yes, so I think a lot of the people -- what you've stated already, I think some of the information was provided about validity, which looked good to me. I looked at it yesterday. Especially regarding the

circularity of the comparison data with the sample of patients that was essentially part of the larger sample, which the measure measures, right.

Then you went on and provided some new analysis, comparing this measure to a different measure again, so maybe you want to explain that. From what I saw, it seemed satisfactory to me as a solution for this circularity problem.

MR. ROMAN: Sure. So we recognize the comment that was made about it kind of being circular in some ways, or one population being a subset of the other, and the target of the measure exactly the same. So we went back and we compared the blood pressure measures to two of our diabetes Hb1c measures instead.

And the results all had moderate to strong correlation so for looking at the blood pressure measure comparing this to our Hb1c control measure where we had Hb1c values of less than an eighth, there was a .81 correlation for the poor control measure --- I'm sorry, I just

said that backwards.

MS. BARTON: This is commercial.

MR. ROMAN: Oh, sorry.

MS. BARTON: And so then the poor control measure is negatively correlated, which you would expect. So I take it, .824.

MR. ROMAN: Right. And in Med -- so in the Medicare line, the same thing with the good control it was .5, and with the poor control it was .57 and in Medicaid it was .79 for the control measure and negative .8 for the poor control measure.

CO-CHAIR NERENZ: Larry?

MEMBER GLANCE: So I appreciate the comments that you all are making and there certainly may be some evidence to support construct validity but I think that in terms of the utilization of this measure, we all recognize that some hypertensives are more difficult to control than others, and there can be a big difference between somebody who is requiring one anti-hypertensive and somebody else who is

requiring three anti-hypertensives to control their blood pressure.

And this measure does not recognize the difference between those types of patients, so I think if you're going to look at the outcomes of primary care doctors in terms of their ability to take care of their patients, you absolutely have to have risk-adjustment. To my way of thinking, this measure needs to go back and needs a little bit of work.

I think it's a very important measure, I think it's an important outcome, but I don't think it should be used to measure the performance of physicians and other providers without some form of risk adjustment.

CO-CHAIR NERENZ: Larry, just a question clarifying question that I have to bring up. This is a plan-level measure only, right? Okay. I understand, but I just ---

MEMBER GLANCE: No, thank you.

CO-CHAIR NERENZ: Okay, Daniel again?

MEMBER DEUTSCHER: Yes, just again, a

similar comment which was the second comment, the main comment that I had was about risk adjustment, so it also comes back to the general discussions we've been having about risk adjustment and the voice that we need to raise here or later on in the process in extending communities, so I guess that there's a question on that as well.

But you mentioned, I think that you did not risk adjust, you mentioned, I think, social risk factors that you did not want to adjust for in order to not over-adjust, but I think there's a wide range between no risk adjustment and over-adjustment, and I think there's still things to discuss so I join the comments here about looking more deeply into those factors that might be important to adjust for so to not misclassify too many providers for this measure.

CO-CHAIR NERENZ: Any other comments? Okay, Jeffrey.

MEMBER GEPPERT: Sort of in a similar

vein, you're looking for, in terms of the characterization of the patients, I'd like to see us delve more deeply into what the quality construct of this measure actually is. What are the behaviors that these plans are engaging in that are resulting in the outcome that we're measuring.

Because then I think you can actually, if you do sort of a correlation like this, where you can tie it more specifically to actual behaviors, then it's a lot more compelling than to just do the correlation without any justification or explanation for why we think these are correlated, and in fact provide some empirical justification for why you think those things are correlated.

We used to talk about creating flow charts and things, we don't really do that so much anymore, but I think in these cases, especially where you don't have explicit measures of quality construct, having a pretty detailed flow chart of why one health plan performs better

than another, having some understanding of why that is would be helpful.

MR. REHM: Thanks for that. So this measure is probably by its number, 0018, one of our oldest measures. The leverage that a health plan can pull to ensure over a long period of time, this measure's probably been in place 25 years or so, are many and they are varied.

And one of the things that is the way of the operating premise, which is in a sort of ecumenical way, there is the target, you figure out what resources you can commit to it, you figure out what the different things are that would be effective in the communities you serve. The primary one is to understand your members. If you don't understand your members, you're not going to make much traction on any of the interventions.

There is no one intervention that is wildly successful across all health plans. I've worked with health plans and depending on your angle of attack, your approach to that whole

population of health management of chronic disease could be quite different.

More than likely, what works for diabetes blood pressure is probably going to be working for hypertension. Not always, but there's similarities, which is why we chose that measure to help support your review. You were looking for another level of analogous, so that's why we responded. I think, in principle, I think that's kind of where we're at.

And I just add one other thought. When we've taken this to the steering, standing committees, five, six times now, one of the things they like, one of the things that our own governance likes, we have three panels that meet on this particular measure, is its broad use to improve population health. And not getting into, if you will, the credible detailed clinical guidelines and how to measure, absolutely bury that, you know, patient profile by patient profile.

Do we need measures to do that?

Absolutely. But from a population health, if you think about health plans that we're trying to hold them accountable for, at a national level, by having benchmark data that helps 450 commercial plans compare themselves.

We do think there's value in that. We also know that the needle's moving on it. That's kind of where our perspective has evolved. Your comments are all forward thinking, and I look forward to addressing them. Thank you.

CO-CHAIR NERENZ: Sherrie?

MEMBER KAPLAN: The age of the patient, I'm not a clinician so I'm way out on a limb here, but it strikes me that plans that serve the diagnosis of hypertension in an 18-year-old in terms of its ideology are wildly different from the diagnosis of, you know, in a 75-year-old.

So do you think, is the argument for not risk-adjusting the plans that serve 18-year-olds are more homogenous and plans that serve 65-year-olds are more homogenous, and adjustment for plan type is enough? Or do you think that is the

ideology so mixed that age is one of those things that leaps out at you apart from diagnoses and other things, is an easy one for you guys to deal with but you didn't?

MS. BARTON: So age is interesting, because I thought that you might say a health plan isn't going to take care of the 18-year-old long enough to see the stroke that's going to happen ten years down the road?

MEMBER KAPLAN: Well, that.

MS. BARTON: And certainly that's true. That's why we think that it's important to hold everybody accountable for blood pressure control, because while there is going to be a handful compared to the number of hypertensives who are 60, there's only going to be a handful who are 18. But the importance of controlling their blood pressure so that they don't have a stroke when they're 40, is extreme.

And so having the plan be accountable for that understanding that they're accountable because there's a measure, somebody's looking to

check that they're doing this right. I think that that's actually an important thing to have, not to say that there should be some kind of, you know, we have to figure out how many of the 18-year-olds have a pheochromocytoma and how many of them are this or that ---

MEMBER KAPLAN: And renal disease or whatever.

MS. BARTON: Yeah, that that would be getting more in the weeds than a national benchmark would want to, given that you're comparing plans that have thousands of members to each other.

MEMBER KAPLAN: Can I ask one follow up question? So a 75/80-year-old after a blood pressure of 140/90, right, and therefore would technically, or 145, if it was a digital thing, over 95, or 93, who's on at least one medication and on a bazillion others and you're fighting to keep everything else under control, you're still being held accountable for that 85-year-old's wacko hypertension. Your argument then is that if

there's not enough of those, or if there's too many of them, you're messing up?

MS. BARTON: Well, the measure only tracks to up to 85. So there's nobody over 85. And, you know, the problem of measures that use a threshold, I'm not going to make a big case for measures that use a threshold because I think that the future of measures that use electronic data, that use digital insights, will be able to do much more, be much more powerful. And I wish that we were there today, and we're not there today.

A measure that says, well, how many of your people with a blood pressure of 190 did you drop by 20 points systolic? That would be a good gold star to have, for clinicians. So for today, what we have, as Bob mentioned, this measure's been around for a long time. It's a measure that holds health plans accountable so they can be compared one to the other on a standard, using a standard definition.

CO-CHAIR NERENZ: Okay. I see three

name tags up. This is now opening to the full panel, I'm going to go Jack, Sean, Eugene and we'll see who else gets on.

MEMBER NEEDLEMAN: Since I'm not quite sure I was heard earlier, I'm going to be very aggressive here, and I'm speaking only for myself, but I've known the NCQA measures from long before I became active on any of the NQA activities. And I deeply appreciate the general philosophy that the plan should be held responsible for figuring out strategies for getting everybody's health under control. I really do. I also appreciate from that perspective the no excuses approach.

On the other hand, risk adjustment is an integral part of any, virtually every outcome measure this committee has looked at and endorsed, it is a standard part of the process, and as I said earlier, there's also an element in the NQF stand which is convenience, in that NQF does not have to go to the individual plans and ask for more analysis than the baseline rates by

asking them for risk adjustment.

And frankly, risk adjustment's going to remain a major element of our assessment of whether the measures get the relative rankings right, get the absolute values across different providers right, and if NCQA wants to continue to come in here without a risk adjustment because of the argument all plans should be held responsible across the board, I think you're going to have to show us that you've done the risk adjustment makes de minimis difference in the rankings of the plans.

Or you're going to have to show us that the underlying factors in a risk adjustment, age distribution, illness distributions, are similar enough across your plan that even if we did risk adjust at the individual patient level, we wouldn't see it at plan level because the factors that are in the risk adjustment are similar enough across the plans.

But we really need to see some effort to look at the effect of risk adjustment in these

measures to be comfortable with your general approach of not requiring it of the plans.

CO-CHAIR NERENZ: Clearly stated. Thank you. Sean?

MEMBER O-BRIEN: Dr. Needleman, very glad you make your argument for risk adjustment very forcefully articulately is no longer needed. I was just curious, is this measure classified in a way that makes it not required to be adjusted, or are we assessing whether their rationale for not providing a risk adjustment is an acceptable rationale?

MS. JOHNSON: Yeah, this is an intermediate clinical outcome so it does fall under that outcome nomenclature, so we would like to have the rationale as to why. And that could be conceptual or it could be empirical.

CO-CHAIR NERENZ: Well, for example, somebody could say, I'd love to do it but I can't get the right data on this. Sorry, I just can't do it. There are a number of approaches, but we just have to judge, for those of you who are

voting, is this an acceptable rationale. Gene?

MEMBER NUNCIO: I want to call your attention to the table you provided, 2E4.2, which looks at the three different types of plans, the commercial plan, the Medicaid plan, and the Medicare plan.

Again, I was wondering who is holding these plans responsible? I mean, what's the impact of reporting that commercial has a score of 54 and Medicare has a plan of 69? That's number one. So who's your audience is going to make a difference here.

The other thing I wanted to call attention to was there seems to be a fair amount of overlap in terms of the percentile ranges, or the percentile values, in terms of performance. So that the 90th percentile for the commercial is higher than Medicaid. I would see this perhaps more valuable to look at distribution within plans as opposed to across the three types of plans that you're dealing with.

That is, I mean, when you look at

commercial and the tenth percentile is 9, and the 90th is 71, that's a huge difference. Clearly something could be done there. Again, for Medicare it's 57 versus 80. You know, 57 starts at the 50th percentile of commercial. So I'm wondering, you know, who's going to make these plans make a change? Where's risk level?

And number two, would you consider looking within plans for your measure as opposed to across plans?

MS. BARTON: So NCQA's role as a creditor is to ask health plans who give credit to report quality measures to us. We release them on our website. We have a quality compass where people can come and look. We also release every year ratings of health plans. And if you follow ANCQA on Twitter, you'll see that there's currently a spate of conversations about which states have the most highly-ranked plans.

You might be familiar with something that CMS runs, called the Medicare Stars Program. And that has been responsible, I would say that

that is what is responsible for the 57 tenth percentile rating among Medicare plans because they pay plans according to their performance on these measures.

If you know of an entity that could pay commercial plans to improve their quality, I would be delighted for you to institute that but right now CMS is the only engine, the best lever, for getting plans to pay attention to quality improvement. And there has been an incredible result to that attention.

MEMBER NUCCIO: I think we all in this room believe that you get what you measure. It clearly is evidence that that might be the case for Medicare, might be good for Medicaid.

Are you looking at all within programs? Are there others that you're reporting domain, your reporting platform?

CO-CHAIR NERENZ: Gene, what do you mean by program?

MEMBER NUCCIO: I mean within a plan, your rate, ranking or rating or whatever,

reporting to the world about commercial plans and which ones have poor performance versus higher performance, and you're doing it based on that sort of table I mentioned.

MR. REHM: Right. So in general, I just picked, I'll pick ACME Health Plan, 38 states, in commercial, and ACME, that's a part of the reporting where it's actually happening out there with 38 states represented here.

And so those 39 ACMEs get to see that data, because that's how we parse it out. We don't call it ACME, we just want data points. We have 38 ACMEs. And they range. They range all over the place.

So that's, you know, this is back to my original premise, which is you look at your own data, and trust me, every health plan has its own data shop that it's grinding this stuff out all the time, probably a very rapid cycle if they care about quality. There's some plans that simply don't. They bills, that's what they do. And shame on them, but that's just an

advertisement.

So that's a little bit about how we construct this. And I think that, from an accountability model, the more people who are in the mix then clearly the more power you have. It's just our accreditation plans could say, you know what? The bar for NCQA is too bloody high, we're out of here. We're going to go to someone else who looks to see if we have policies and procedures and we'll accredit you that way.

That's an option. So in some ways we're trying to add measurement as the guiding force. People don't vary quality in terms of their compliance with the standards, the structural measures. They generally all do very well, because they know exactly what to do. Influencing your members' behavior, influencing a community's health, influencing physician behavior, that's a tougher road to hoe. So that's why we like this spotlight.

CO-CHAIR NERENZ: Let's just make sure this stays on decision and comments. I'm starting

to wonder here, but go ahead.

MEMBER NUCCIO: Just a quick comment. I mean, these data are the long risk-adjusted data. I do support my colleagues in their desire for making sure that we're comparing apples to apples.

MR. REHM: Comparing health plan to health plan. So, yeah.

MS. BARTON: Well, and I'm sure that your colleagues in Medicaid health plans would much prefer to be compared to other Medicaid health plans. They see that as the fairest way to rank them, to rate them.

CO-CHAIR NERENZ: All right, I think points have been made, I don't see any other tent cards up, so let us call the question on this one. Subgroup 2, I think, are still the voters for real. The rest of us, the Survey Monkey.

MS. OGUNBEMI: We are now voting on Measure 0018, overall relating of validity, and your voting is open.

Okay, we're at zero percent high, 67

percent moderate, 33 percent low and zero percent insufficient. The measure passes validity. The counts are zero votes high, four votes moderate, two votes low and zero votes insufficient.

CO-CHAIR NERENZ: All right, thank you. All right, we now, last one for Subgroup 2, we are now moving to Measure 3534, 30 Day All-cause Risks Standardized Mortality Odds Ratio. We're only missing two necessary hyphens up there; sorry, that's my pet peeve.

(Off microphone comments.)

CO-CHAIR NERENZ: Mortality Odds Ratio, and we will have Karen do this one up for us.

MS. JOHNSON: Yeah. How well you, that's a long title to get your tongue around. So this measure estimates risk-standardized 30 Day All-cause Risks Standardized Mortality Odds Ratio following Transcatheter Aortic Valve Replacement (TAVR). It uses clinical data available in the SCS ACC TVT registry for risk adjustment.

So when they developed the measure, it used site-reported 30 day follow up data in that

same registry.

So this is an outcome measure, again using registry data. The level of analysis is a facility level. It is risk-adjusted with 41 factors. So when it came to our subgroup votes, it did not pass reliability. And for validity, there was not consensus reached on this measure. This is a new measure, so it hasn't been seen before.

And, let's see. For reliability, I'll just give you the real basic stuff and then we can go to Alex to get more into the concerns. For reliability, the developer-assessed inter-rater reliability using data from 40 records randomly selected from four randomly selected facilities, I hope I have that down right, so presumably that was ten records per facility but I might not have that exactly correct.

For validity, to do data element validation we did two different methods. They looked at record eligibility assessment, so looked at six hospitals. And basically looked at

all of their TABT and mitral cases within a specified time frame and then compared that to the registry to make sure that none of the records were being missed. And then they looked at 40 hospitals that had at least ten cases randomly selected for audit.

So they looked at ten baseline, ten follow up cases, randomly selected for abstraction, so therefore ended up with 400 baseline records, 400 30-day records and quite a few fewer, 289 one-year records. So they did a prevalence-adjusted, bias-adjusted kappa statistic.

The key concerns around the validity analysis included exclusion of more than 50 percent all hospitals and patients due to missing data, relatively low values for that adjusted kappa for two of the values, lack of data element testing for most of the variables and a relatively small testing sample that may or may not be representative of hospitals and patients included in the measure. Those were some of the

main concerns. I think we hand it to Alex. Is that right?

MEMBER SOX-HARRIS: I'm going to briefly focus on three out of the four main issues as I saw them. The first being the eligibility criteria for participation in the measure. They dropped a lot of sites for eligibility, then there is what is considered a critical element to be subjected to both reliability and validity item level testing.

The item level reliability results, I think we're off the hook to discuss some of the concerning aspects of that methodology, because if I understood earlier, they do item-level validity analysis, then we don't need to consider the reliability analysis, which is nice, so.

So I want to thank the developers for responding to some of our initial comments. They provided that data and rationale that were very helpful.

To talk about the eligibility requirement for participation in the measure,

there are 480-some entities that participate in the registry that the measure is based on. In order to participate in the specific measure, developers have said that all critical elements, all variables, need to be at least 90 percent present in the data set for entities to participate.

The effect this has is it drops out over, I think roughly half of the entities. And so this is, they've provided a nice table, so if you scroll down to the supplemental material they've provided, I think it's figure -- let's see if I can find this quickly -- figure B-1 or 46, it shows the tradeoff between if you have different thresholds of allowable missingness, what the consequences are in terms of how many sites you can get included.

I think they present this data for our consideration in terms of whether their judgment of 90 percent is the correct tradeoff between saying something potentially more accurate on a sample of the population of entities versus

including more entities and having to deal with more missing data. So that's the eligibility question.

The critical element question, so their initial submission identified of the 40-some risk adjustment variables, they identified five of them as being critical. When I was reviewing this, I'm a newcomer to this group, by saying, aren't all the risk-adjustment variables critical? I mean, if you have serious reliability problems or validity problems on any of the risk adjustment variables, that seems to be important.

So that might be a clarifying thing for NQF, to say when we say, or when NQF says critical elements, does it really mean all the elements that you need to make a measure, or is it some subset of one of the criteria to make that distinction?

The developers kindly went back and did 21 or 24 of the variables, subjected them to reliability or validity testing. That's not all of the variables, but it's a lot more than five,

so I found that helpful.

Last point is on the item validity testing. So again, if you go to their supplemental table, what you'll see is they've done -- so one of the issues was essentially audit, where they took data from the registry, sample registry entries, and then went into the component systems and looked at their records to do an audit to see if the registry was accurate. So it's strong item level validity analysis.

One of the questions was, were the selected entities and patients similar to the non-selected ones? They provide a nice analysis to show that there's nothing really concerning about the selection at that level.

When they went through and did the prevalence-adjusted kappas for the different elements, most of them looked excellent although there are two that have concerningly low values, including one that may be essential, which is the date of death, since it's a 30-day mortality measure, not having a good measure on date of

death to be problematic.

I'll take a breath and allow my colleagues on the subcommittee to chime in.

CO-CHAIR NERENZ: Just before I go there, could I just ask you a clarifying question? It's on this point exactly, date of death, because I was looking at this and saying holy smokes, how can you have a measure if you don't know the date of death? But actually, I think I have an answer to my own question. I want to know if it's the right answer.

In a registry like this, let's say you do a case review in 30 days. You can record in the registry yes/no whether death has occurred, and you can be quite accurate of that even if you didn't know for sure if it was day 25, 24, 23, and so it seems to me theoretically possible that you could be very, very accurate on yes/no, death has occurred, and then you can spin this measure even if you don't know the exact date.

So is that what's going on here?

MEMBER SOX-HARRIS: I think you're

right, but it's hard to check if yes/no is correct unless you know the date. That's the issue.

CO-CHAIR NERENZ: Maybe. Well --

MEMBER GLANCE: You know, it's funny, when I was reviewing this, I didn't rightfully understand why they went about presenting it the way they did. So, in other words, why did they focus on the data elements, when in instance of data registry set, this is high-quality data. This is STSs. They know how to do really, really great measures. Why didn't they just go through and do the usual stuff, which is to basically present the reliability and the validity of the score measured?

I think that part of our job, honestly, as a panel is just to provide feedback. Just like a reviewer would for a peer-review journal. Personally, I think the two things I would suggest would be go back and just do what you normally do which is look at the score, measure of validity and reliability.

The second thing, and I think this is probably the more substantive comment, and I think you made this comment, is that if you're going to look at, you're going to use a measure to do bench marking, to do accountability, you can't exclude 50 percent of the patients.

I think they did this because they wanted to include these really cool functional measures in their model. And they really are kind of cool, and they may be somewhat important in terms of risk predictions but you know, you can make the same argument for virtually any of the other STS or non-STS measures. But the reality is they were missing them in 50 percent of the hospitals. You can't use it if it's missing in so many hospitals.

So I guess our feedback to STS would be, I think it would be really straightforward because I'm sure they have these calculations already done, would be to look at the measure and look at reliability and validity at the score level, number one; and number two, reestimate

your model with these two risk factors that are missing in virtually half the patients.

But I think at the end of the day, it's hard to pass this measure based on what was presented to us in terms of validity.

CO-CHAIR NERENZ: And also, just as a comment to the group, I'm looking up here and I see the main problem actually being in reliability. That's just a flat-out fail. Yeah, I think. And then validity's a consensus in a breeze. So does anybody, particularly in light of this new information, want to argue for rethinking of some of those reliability ratings? Because that's just a straight fail, right there.

Yeah, Alex?

MEMBER SOX-HARRIS: I had some real concerns with their reliability item-level methodology. But as I said at the outset, I don't think we need to judge that if their item-level validity analysis is convincing, is that right? I think that drove some of the low reliability ratings.

CO-CHAIR NERENZ: Okay, so let's make sure we clarify that. Because if somebody rated reliability insufficient, but the rules of the game said they didn't even have to do it, that would be, I mean, I guess you could still vote it, but that's --- what are we doing here?

MS. JOHNSON: Yeah. So what you need to do, you can take their updated reliability information if you want to, or you can take their updated data element validity information and apply that.

So since they did do data element validation, you can take your rating for that and apply it to reliability. So, and I think as Alex stated, they originally only gave you five variables. They gave you many more after that, so take into account you have more variables than you did, I don't remember if the sample size changed or if it's the same sample size, I don't recall. It's about 400. Yeah.

So am I being clear? You guys understand? I suspect that the low numbers on the

reliability the first time around had to do a lot with having only five elements. But that was my guess. But you can take what they -- because they updated both, right? They updated reliability and validity, if I remember correctly.

CO-CHAIR NERENZ: All right. Not seeing any nametags. Do we have somebody from ACC on the phone?

MEASURE DEVELOPER: This is Susan. I'm ACC staff.

CO-CHAIR NERENZ: All right, Susan, I assume you've been listening in. Any comment, response that you'd like to make so far?

MEASURE DEVELOPER: Well, I think we recognize some of the things that you're expressing concerns about with this measure. It was one of our first toes in the water for something we're attempting to publicly report. We had talked so many times about the, it's -- all the variables are very complete.

The mortality date is missing. Sometimes it's a rare occasion, and there is

something that attached to our response, actually. We have done a check with CMS data to see if there was any gaining or differences with CMS, so that, we really never saw any mismatches, it's like 20-something percent of the time, over several years.

But the, first of all, the inclusion criteria, the having the KCCQ, it's something we kind of felt that sites would man up and start complying. They have improved over time as we just started publishing this measure in our outcomes reports not long ago. We feel like it's important.

KCCQ is part of the coverage decision, so the physicians really, and Dave Shavian has worked with us for this and he said, people raise the bar for STS, when STS raised the bar people kind of stepped up to it. Everybody felt like sites would do that for us. So that's one thing.

And then the testing, we just decided not to submit this data in the first round and we would, if this measure was endorsed we would

submit it a couple of years from now. Because the way I understand this, and I'm not, I haven't really worked with this, but the way I understand it is that it's not something required. If you submit your audit and your interreliability, you're not required for the criteria to submit the testing.

So we do typically submit that at STS and ACC for other measures.

CO-CHAIR NERENZ: All right. Thank you. I guess I'll defer to staff if they want to respond immediately to that last, it's really sort of a statement about the rules of the game.

MS. JOHNSON: Yes, she's correct. This is what we call a health outcome measure, so we ask for either data element testing or score level, and currently we do not require both. What you're hearing is Larry saying, golly, it would have been nice if you had given us both, but he knows that he can't take it down because you didn't. But he would really like to see it in the future.

MEASURE DEVELOPER: Yeah, and we know that and we know, I mean one thing we looked at is there are lower volume sites have lower, have less select reliability in their measure, and if we're publicly reporting it, that's going to have to be something that we need to talk through. Because we don't want a lower-volume hospital to be reported as not as reliable.

So the reliability, there's really low-volume hospitals and there's hospitals with really high volumes. So that was one of the challenges with that testing.

CO-CHAIR NERENZ: Dan?

MEMBER DEUTSCHER: Is there any way the developers could comment on the discussion on the percent of completeness threshold that they set? I found the additional table provided very interesting. From 440, 450 sites, applying the current threshold, only 188 are included, which is a comment that was brought up here before. So, any comments on this threshold and the tradeoff maybe going to a slightly lower completeness

threshold?

I was looking, for example, at the 70 percent or above, seemed to have resolved in a completeness of the three columns here, all over 90 percent. Just as an example. But maybe a comment on that would be helpful.

MEASURE DEVELOPER: Well, I mean, I think one thing is that the criteria, is that you perform the 30-day follow up and the statement. You have to perform all three of those on 90 percent of your patients to get your outcomes.

That table that's in your packet is from model development, and in the application of the testing, I think I showed from model development until our most recently published report how that has increased over time, the number of sites included, because they are really trying to work on that.

But from what I know about those two variables is we could drop the variables, and we wanted to see if we can keep them. But you can't, if a site has incomplete KCCQ or incomplete Five

Meter Walk, you impute a median score. That might be good or it might be bad for a patient. It may make a flight's risk adjustment artificially higher or lower.

Those are two tools that it's just more difficult to impute the median, the median score, which is what you would do on a survey with a Five Meter Walk if the patient didn't have this assessment. So that is kind of challenging in just lowering the inclusion criteria but imputing the missing patients to a median score.

MEMBER KAPLAN: I have a question about NQS position on imputation because, you know, to the extent that you have more opportunities for missing data, you have more missing data. So multiple imputation is sometimes indicated, sometimes not.

For example, if it's associated more closely with a dependent variable of whatever quality measure, you might not might want to impute rates, for example. It depends on circumstances. What is your position on

imputation, and what's called for in certain circumstances, or does NQF have a policy on this?

MEASURE DEVELOPER: Actually, NQF to my knowledge has never made a policy on imputation, so we don't have any statements at all. Or if we do, it's in the back of time, before I started. As far as I know, we have not made any statements about imputation.

MEMBER KAPLAN: Can we put that in the parking lot for tomorrow?

MEASURE DEVELOPER: Yeah, maybe you guys can make some statements.

CO-CHAIR NERENZ: I agree. It's, I was maybe going to use a little bit terminology and not so much the imputation, but what do you do in real life, but in measurement as we're judging it. If you've got a complex risk-adjustment model but a lot of data elements are frequently missing, do you just exclude all cases that have any variables missing, which is what my own statisticians do. It drives me crazy.

Or, you know, do you come up with some

sort of way in which you can use what you have and argue that that's better than nothing. I don't know the answer to that. Somebody has to know more than I do, but this is something that we probably shouldn't speak to.

MS. JOHNSON: Or you do a sensitivity analysis to your conclusions, based upon different imputation strategies.

MEMBER GLANCE: One thing that we've done in our group is basically you take a sample where you have all the complete data, and then you just basically get rid of the complete data on a subset of the data, then you impute that missing data. Then you see how well your impute data is working.

MS. JOHNSON: You do the sensitivity analysis.

MEMBER GLANCE: Yes, exactly. You can do that.

MS. JOHNSON: But I didn't know if they had a policy on it.

CO-CHAIR NERENZ: Interesting. Okay. So

this sounds like a, when we're talking about how to make the world a better place tomorrow morning, or early afternoon.

Okay. Let's, any other comments from anyone?

MEMBER GLANCE: I was just going to say that, I will not take down this measure. I saw that look. I really think we need to talk about this. I think the idea that a measure can pass just because somebody's looked at a hundred copies --

MEMBER ROMANO: Can you speak louder? We can't hear you.

MEMBER GLANCE: You can't?

If you can just --

MEMBER GLANCE: I was just going to say that, I think we really need to talk tomorrow about whether or not it's acceptable to only look at the data quality. Because the idea that someone can look at hundred charts and assess validity and reliability at the data element level and that's enough to get any endorsement,

that to me kind of is a problem.

CO-CHAIR NERENZ: Now let me just check the question. For those of you who are going to vote for Rio, are you okay with these convoluted rules of the game, and validity counts for reliability and the measures disclosed are required? I just want to make sure everybody is okay on this, because it's going to matter here.

MEMBER GLANCE: We're supposed to impute our reliability, I --

(Laughter.)

MEMBER HYDER: Is this like a jury trial? Are you supposed to excuse yourself in front of the jury?

MEMBER GLANCE: No, we're just going to nail it. All right, I just want to make sure everybody's okay, those of you on the subject.

MEMBER HYDER: I'm okay with it, but there's a new agenda item called "boil the ocean," that I would definitely engage.

CO-CHAIR NERENZ: Okay, all right. We'll get to this.

MS. JOHNSON: And just for clarification, they did give us updated reliability information, so you can use that. We're not saying you have to use the validity data element, validation, right? So you can look at what they gave you with reliability, that's, you know, you can use that. If that's not enough to pass them, then go ahead and look at their data element validation and see if that is enough in your estimation to give them a passing rate. Does that make sense?

MEMBER FABIAN: In order to be a true, unbiased evaluator, you spent, I probably wasn't paying as much attention as I should have been, because I had my preconceived, you spent 90 seconds reviewing that updated reliability and validity information on the data? Because I don't have it in front of me for some reason.

MS. JOHNSON: Yes, it's --- you might know it faster and better than me.

MEASURE DEVELOPER: So it's in there, the supplemental stuff they sent. And let's see,

discussion guide, page 95 and beyond. Is that it?

MS. JOHNSON: Yes. It's possibly 99.

MEASURE DEVELOPER: It's --

CO-CHAIR NERENZ: 95?

MS. JOHNSON: 95, yes.

CO-CHAIR NERENZ: Okay.

UNIDENTIFIED SPEAKER: 95.

MS. JOHNSON: 95,96, oh, yes. I wrote that down in the --

CO-CHAIR NERENZ: Okay. I'll look at it right here. Okay.

MEASURE DEVELOPER: I think this validity guidance stuff starts on 100, page 100.

MS. JOHNSON: Is that clear? It's --

MEASURE DEVELOPER: Yes, no, I see it, thank you. Thanks for, you know, --

PARTICIPANT: It's page 98. It's the beginning of validity.

MS. JOHNSON: So let's start with page 95 for reliability. Start with page 98 for

updated validity.

CO-CHAIR NERENZ: Let's pause just for a couple of minutes, let folks look into that if they want too. I'm sure they'll get it right. Let us move now on to voting. As we've been, Subgroup 2 does it for real, the rest of us on SurveyMonkey.

MS. OGUNBEMI: We are now voting on the reliability for Measure 3534. Your voting is open. The options are high for moderate, low and insufficient.

So the voting is closed. The results for Measure 3534 are resounding moderate. All six participants voted moderate. So we have zero high, low, and insufficient, six votes moderate, 100 percent moderate votes. So the measure passes on reliability.

And we are now voting on validity Measure 3534, your voting is open. Your options are, high, moderate, low and insufficient. Voting is closed. The results are: 0 percent high, 83 percent moderate, 17 percent low, and 0

percent insufficient. That is, zero votes high, five votes moderate, one vote low, and zero votes insufficient. Measure 3534 passes on validity. And we just move to Subgroup 3.

CO-CHAIR NERENZ: All right. Thank you again, everyone. Moving right along, we are now basically on time, I think. We have 35 minutes from here to our scheduled next break point for the first measure out of Subgroup 3. So Subgroups 1 and 2 can relax. This will be Measure 3478. This is surgical treatment complications for localized prostate cancer. The issue -- issues plural are that it was consensus not reached on both reliability and validity. So we have to see if we can sort that out. And Ashlie.

MS. WILBON: Are the -- is the developer in the room? Okay. Please come have a seat at the bench.

MS. ANDEL: Yes. I'm Corinna Andel. I represent the ADCC Quality Committee.

CO-CHAIR NERENZ: Welcome.

MS. ANDEL: Thank you.

MS. WILBON: You're welcome, Corinna. So just a quick note, Subgroup 3 is John Bott, Marybeth Farquhar, Paul -- I'm sorry, he's not here -- Dave Nerenz, Eugene Nuccio, and Ron Walters. So you guys are on deck starting now for Subgroup 3, and we're looking at Measure 3478.

So this is a new measure. The description is that the measure analyzes hospital facility level variation in patient relevant outcomes during the year after prostate directed surgery. The measure uses claims to identify urinary incontinence and/or erectile dysfunction among patients undergoing localized prostate cancer surgery. So I'm going to read through the numerator and denominator because it was -- it's a bit tricky. So I think that's important in kind of understanding how the measure works. So we'll have the developer clarify if I miss anything, but we'll just walk through here, the numerator and the denominator here.

So we'll start with the denominator. So the denominator is Medicare fee for service beneficiaries, 66 years of age or greater, with prostate cancer who receive prostate cancer surgery -- so open or closed, if you will -- after prostate diagnosis and survived at least one year after the surgery.

The numerator is -- so the numerator generates a hospital level score of incontinence and/or erectile dysfunction during the year after versus the year before prostate surgery. So each individual patient's change and the number of claim stays for incontinence and/or erectile dysfunction calculated in the year after versus the year before the surgery.

The patient's scores are then truncated and then re-scaled from I think negative five to 10 to scale of zero to 100. The hospital score is the mean difference in days among all attributed patients. So that's for the hospital. The claims for incontinence and/or erectile dysfunction during the year after versus

the year before. And it uses claims for both inpatient and outpatient encounters. So I just wanted to give that kind of brief overview so folks have a sense of what that measure is capturing. Move back to the discussion guide to summarize some of what the -- some other of the preliminary analysis votes.

Again, this is at the facility level, and it's not risk-adjusted. The ratings for reliability, again were, because consensus were not reached. We had three moderate, one low and one insufficient. Score level reliability was demonstrated using a split-half affirmed with Pearson correlations for the minimum sample size of 10, n equals 10, which for 233 hospitals of correlation was about 0.65. For the largest sample size, which was about, which was n of 80, the correlation was about 0.89. I also wanted to note that again, I think this has come up with some other measures, that we only require score or data element reliability testing in order to meet our requirements. I think that came up in

some of the reviewers' comments about their concerns about the measure, so I just wanted to clarify that essentially going forward, so we don't kind of conflate those requirements with what you'd like to see. And there was also some concerns over the lack of clarity of this specification, some confusion about how the attribution works, some of the time frames about how the scoring methodology was implemented, if the winterization method is for kind of truncating the scores if you will. I also wanted to note, there was a few reviewers who noted about the age of the data. So they used data that was tested from 2009 to 2013 using ICD-9 codes. And we did have kind of prior approval with this developer that they could submit testing using ICD-9 codes based on a waiver that we granted due to some limitations with data that they identified. So if the measure is endorsed and they come back for maintenance, they will be required to submit ICD data, testing on ICD data at that time.

For validity, it was also consensus not reached. They did do an assessment on validity doing a systematic assessment of face validity. Their assessment does meet NQF requirements for face validity and is acceptable for demonstrating validity for new measure submissions. So, again, for this concern, I think that some reviewers noted that we wanted to see other types of validity demonstrated for a new measure. The first time in the door, we do accept face validity, although you still should also weigh other threats to validity. But in terms of the testing requirement, they have met that with their face validity demonstration.

Other concerns with validity were around the lack of risk adjustments on the analysis of the exclusions and missing data. So with that, I will hand it over to the lead discussant which is John.

MR. BOTT: So Ashlie covered things pretty well. I'll just elaborate a little bit on a couple of things that stuck out to me and that

other folks on the group hit on as well that are here. So I was the one who voted low regarding reliability, and it was because of the ambiguity I thought in the specification. So while I had a number of them, I'll just point out a couple. One was this concept used in the exclusions and the numerator at a minimum, maybe it was used in that elsewhere as well, that the quantity of days, somehow using claims codes to get at quantity of days, that's new to me. I've been in the game of working with client data for a long, long time.

So for example, in the exclusions, there are several exclusions about quantity of days. One example is verbatim, two or more days with code for secondary malignant neoplasm. I don't know how claims codes are used to get a two or more days, I don't understand that, and it's not explained. And numerator also has this concept with number of days where it just says numerator is identified from claims, in parenthesis, days with claims.

So I understand what the numerator events are, but I'm not sure how they're getting that number of days based on claims. I would expect if you're using claims, that you somehow count the number of days in a novel way that should be enumerated more than it is here. Another area is this statement about attribution. After the list of the denominator bullets, it states: "Patients are then attributed to the hospital/facility associated to the claim for the procedure code for prostatectomy." But it doesn't talk about -- it's just silent on how that attribution happens. And it seems like with measures like this where you're looking at an event in the hospital which you would sort of call your in-depth submission for other measures where you're counting numerator events post-discharge, measures oftentimes get into attribution like readmission measures. So things that it seems like it should take into account, but I don't know if it does, because it's silent is: what if the person had a qualifying

denominator procedure performed in another facility? It doesn't.

It's also silent on transferring into that hospital where the procedure is performed, or what if they transfer out to another facility? So it makes me think there -- the attribution is simply the event happened at the hospital. But then the last denominator exclusion states, "patients who could not be attributed to a hospital." So it does make me think that, well it seems like there is some attributional logic if you would state that. But I think that we should be -- we should not be guessing at these types of things. So those were a couple of things that jumped out at me more.

A couple other notes related to reliability that others made. Somebody said there is a potential bias here because the procedures for incontinence are medications for erectile dysfunction. There could be a bias that I think the person is saying, and if they want to jump in, those aren't accounted for. Maybe the person

is suggesting that these should be exclusions then to level the playing field.

Also, I did have a somewhat smaller question related to the winsorization, which I won't get into because I noted my high-profile ones. But somebody else noted that the way the scores are transformed from the range of a negative five to 10 get transformed from I think zero to 100. And somebody stated under reliability that that transformation may define differences perhaps more than is reasonable. So that's the reliability.

A couple quick things on validity, but I think I'm risking repeating what Ashlie may have said. One of the concerns expressed by somebody was there's other conditions that could impact incontinence, suggesting that that could have been tested for in risk adjustment. The person noted stroke, Parkinson's, diabetes.

I was surprised that those all seem logical, but the measure developer cited they did look at an ICHOM report, they did a lit review,

they had a tech panel, so I am surprised those didn't come up and weren't tested for. So again, the other panelists who I noted, they might want to elaborate on if we're aware of evidence that those are needed and related to risk of incontinence.

Other complications could be having an impact on the rate. It wasn't -- so again, if somebody else wanted to, if somebody else noted that, they might want to jump in. Lastly, there was a concern about: is this measure really expressing meaningful differences to the point where we could really be measuring variation and attribute it to the hospital?

So that's the high-level validity and reliability concerns for the panel. But if anybody wants to jump in with one of the points that they have?

CO-CHAIR NERENZ: Yes, and also I'll just sort of reframe little. Of those on this particular Subgroup, I know that's a pretty long list of things for which we need a response.

Anybody else with challenges, questions, concerns, let's get them all on the table at once. And then I know you're -- I'm watching you taking good notes here.

MEASURE DEVELOPER: I'm taking some notes, I'm going to need a reminder.

CO-CHAIR NERENZ: Okay.

MEMBER NUCCIO: Yes, I'll only, the question about rescaling from the negative five to 10? When you expand that to 100, you're going to create mathematically significant differences that may not be meaningful. If you really put it that way, okay, mathematically different, probably nothing major. That's number one.

Number two, I was confused by the higher equal better in terms of the metric itself, in that it sounded like it, providers that create more E.D after the prostate surgery were rated better, okay. And I don't think that's what you meant. And then it has to do with that, as what I think Ashlie was talking about, the looking at differences. So if there's

something you can do to clarify the language there.

And in terms of the -- I'm a little concerned about the risk adjustment. You did provide some correlations for potential items. And I understand the lack of data issue, that I understand. But of the potential risk factors, you only have the Gleason scales of which is a severity issue -- who was the only one that had any kind of inkling. And so I think I was surprised that your tech did not come up with something more. And I don't -- maybe we can go back and repoll them or something, but I found that strange that they didn't provide you with a laundry list of things that they could made a difference. But I mean, you know, even if it's where the surgery takes place and that was -- John talked about that. How it was possible, or outpatient clinic, or --

CO-CHAIR NERENZ: And I'll just sort of -- I'll do it as if I was responding to John's question, but is actually -- you're going to

clarify whether I got this right. I as a member of this group wasn't too worried myself about this claim state thing. But now I'm interested in your discussion of this. Well, fundamentally, it seemed to me like situations, like for example, if you're worried about the effect of diabetes, for example, it's a comorbidity, and you want to identify comorbid patients, you're looking for two claims with the code some days apart, just to indicate it's not a false positive. And I sort of then extended it to here to say, the fundamental premise seems to be that if you're accepting one of these two problems, the more it shows up on a claim, the worse it is or the more frequently you're getting treated for it. And that is the fundamental concept of the numerator. And it's also --- it sort of runs negative to positive, in that in principle you can be having codes for these prior to surgery, and then it goes away. And you don't -- okay, so at least that's the concept. Now whether we buy it and say that's right, okay, different thing.

But that's at least --

MEASURE DEVELOPER: Well put. Yes.

CO-CHAIR NERENZ: Okay. Bott, so why don't you go ahead and I'll let you --

MEMBER NEEDLEMAN: David? Could I, I'd like to have a little bit more inclination about where the codes are coming from when we say number of days. So there's a claim in October that includes erectile dysfunction, and then there's another claim that the surgery is in November, and then there's another claim in January that includes erectile dysfunction. So you're counting all the days between the surgery, either looking back or looking forward based upon the dates of those claims?

CO-CHAIR NERENZ: No, it's --

MEMBER NEEDLEMAN: That's what I'm not understanding. How are the actual numerators being calculated?

CO-CHAIR NERENZ: Okay. So you've got one before, you've got one after, the change is zero. That's what I thought they were doing.

MEASURE DEVELOPER: Yes.

CO-CHAIR NERENZ: The day interval is meaningless, right?

MEASURE DEVELOPER: Correct.

MEMBER NEEDLEMAN: So it's simply the number of times you have been in for a visit, an outpatient visit, or an inpatient visit, for which this the diagnosis that is primary? Secondary?

MEASURE DEVELOPER: If the claim --- it's the number of days with claims for either of those conditions, either as an outpatient or an inpatient.

MEMBER NEEDLEMAN: Okay. Well, this is what's confusing. If it's the number of claims, then I understand that. But if it's the number of days with claims?

MEASURE DEVELOPER: I'm sorry. It is the number of claimed days --- claims-based days. So the number of days with a claim, either of those conditions.

CO-CHAIR NERENZ: Well, and -- I

shouldn't be in the business of defending, but this is what I'm verifying. I'm not trying to say it's a good event. Let's say you've got one of these problems, you go in in a clinic and you see in sequence three different specialists, just cause. And each one of those generates a claim and a Medicare fee for service. And essentially the idea is that's one visit. That's how the day concept comes in. It only counts once, not three times, but --

MEMBER NEEDLEMAN: But we're not looking at the duration, of the --

MEASURE DEVELOPER: Primary.

MEMBER NEEDLEMAN: -- of the problem, we're looking at the number of days you have gone in for actual diagnosis and treatment.

CO-CHAIR NERENZ: Yes.

MR. BOTT: Yes, I don't mean to stifle conversation with the measure developer, but I did specifically ask this morning if the measure developer had the opportunity to clarify the questions that we posed, primarily I posed, and

failed to -- which the measure developer failed to address these questions in writing. And so it would be in the material that we agreed before the meeting. Karen had said that they will not have the opportunity to allow verbally then to address them in person.

MS. JOHNSON: Sir, I didn't --

MR. BOTT: Well, maybe it wasn't Karen, it was somebody else, but I thought that it was you who said that.

MS. JOHNSON: Yes, they can address verbally. They can't give you anything written today.

MR. BOTT: Okay.

MS. JOHNSON: So apologies if I said it wrong.

CO-CHAIR NERENZ: Go ahead.

MEASURE DEVELOPER: Okay. So in terms of attribution, the hospital attribution, so what we did is we first looked for a claim with a procedure code for prostatectomy either in the outside file or in the MedPAR file. Okay. So we

start with that and then we identify the hospital or the facility that's associated with that claim. And then we take a look at the patients for whom they had claims for a prostatectomy, but only in the carrier file. So we could only identify the condition. But we could not find a hospital where that procedure took place, and we excluded those. And that was in 275 cases out of more than 10,000. But that's how we wanted to determine where the procedure occurred. We wanted to be able to attribute it to a hospital.

The winsorization and the concern about magnifying the differences in the rescaling. So the winsorizing occurred, the cutoff was plus or minus two standard deviations. It was not an arbitrary number, negative 10 or five, it was -- that's what the plus or minus two standard deviation cut off, so that's why we picked that, and that has implications to the rescaling.

In terms of magnifying the difference, our impression was that because we were adding

constants to all of the numbers, it doesn't magnify the difference. Because you're making that change to every single score at the same time.

Let's see, and then in terms of meaningful differences, taking a look at that among the facilities, we do recommend -- and we have provided in the submission form -- that the hospital's claims file score be included so that you can see where the hospitals fall under those sections. And at the very least we can have comparisons among those. So we do note that that would be helpful, and we did detect that they're an issue among the hospitals. We think that this claims-based measure could be useful in that regard in identifying differences between hospitals on either end of the spectrum or across it. Let's see, the risk adjustment, we actually did look at a number of patient variables and hospital factors, and then you looked at what came out and tumor factors and looked at what was significant.

And then we ran the numbers. We adjusted them using the significant patient level factors. We also looked at adjusting the scores of the hospital and patient level factors, compared those adjusted scores to the unadjusted scores, and found that there was a correlation of .95 in both cases. And that led us to the decision that risk adjustment was not necessary and added unnecessary complexity.

Okay, what else here? In terms of how factors such as stroke or diabetes could impact the results, we also did take a look at comorbidities, and those were not significant. That is in the packet under -- in one of the tables. Also, because the patients serve as their own control, if they have that stroke or diabetes or other comorbidity prior to the prostatectomy, they're going to sit the evaluation out. Let's see, what were some of the other questions I have not addressed yet?

MEMBER NUCCIO: I'm sorry to interrupt, but I understand that you truncated

negative five to 10 based on two standard deviations.

MEASURE DEVELOPER: Yes.

MEMBER NUCCIO: That's great. But then how did you convert those values from zero to 100, so that 10 became 100 and negative five became zero?

MEASURE DEVELOPER: Yes, okay. So there were a number of steps that we identified. So what we did was we looked at -- we subtracted the difference after we subtracted the complications from prior to the surgery from after a surgery. And then we winsorized, and then we took those different scores for each patient, subtracted 10 and then multiplied by 10 over -- 100 over 15, and that created the scale from zero to 100. And we wanted to create a scale from zero to 100 because we found that would be more intuitively interpretable for the average person. We know what a score of 100 feels like versus 40, as opposed to looking at negative five to 10. You know, there's no

meaningful anchor for that.

MEMBER NUCCIO: So then proportionately, the ends expanded more than the scores in the middle?

MEASURE DEVELOPER: Well, but you're doing the same thing to all of the scores.

MEMBER NUCCIO: Oh, assuming numerically they change. Okay. I understand I think.

MEMBER ROMANO: Yes, I'm curious about this formulation with subtraction. I think it's fairly self-evident that if someone has no claims before the surgery for one year and then they have one or more claims after the surgery, that something happened, probably, related to the surgery.

But I'm not sure what it means if somebody had, let's say, two claims before the surgery and then they had three or four claims after the surgery. And so the subtraction yields a positive number, but can we really interpret that as a worsening?

So I'm curious if you have a sense of, when this indicator flags positive, how often is it because of this increase in the number versus going from zero to a positive number?

And have you done or are you planning to do anything to look at the validity of, you know, going from two to three or two to four, is that really a worsening from a patient's perspective?

And I say this because I think we all agree that the right way to measure this concept is with a PRO-PF with a patient reported outcome measure. So what you're doing here is a proxy and the question is, how good is it for men who have these problems before and after the surgery?

MEASURE DEVELOPER: Well, I think we would argue that if they had more claims following surgery that something happened to make it worse compared to before, prior to that, prior to having the surgery.

Your point is well taken about collecting PRO data for this kind of thing. And,

in fact, we are working on some PRO development, but that takes a long time and it's going to be a few years before we are able collect the data for such a measure. So --

MEMBER ROMANO: So do most of them flagged positive, are they mostly zero before, or are they mostly have some claims before them and some after?

MEASURE DEVELOPER: I don't have the answer to that, I can't tell you.

MEMBER NEEDLEMAN: Can I follow up? I'm curious to know why you think the number of claims is a better measure than simply the condition is present or absent?

What is it that I need to understand about the clinical presentation of the patient with these conditions that says more days visiting the doctor indicates a worse condition?

MEASURE DEVELOPER: Well, prostate cancer is associated with a number of complications. And it's possible that undergoing surgery, could actually make them worse instead

of better.

MEMBER NEEDLEMAN: Yes, I got that. The question is, why do more claims post-surgery before erectile dysfunction indicate that the condition has gotten worse?

MEASURE DEVELOPER: It's an inference we're making. It's, I mean it's --

MEMBER NEEDLEMAN: I mean it's quite plausible that if somebody has erectile dysfunction, let's say preoperatively, that the surgeon might say, well, let's deal with that after surgery.

And so, they would specifically defer dealing with that problem knowing that the patient's going to have surgery. And then all of a sudden after surgery, you'd see a series of visits for that problem.

And that was part of the plan. So I think that maybe what we're getting at, and it's not entirely clear that the number of claims is good proxy for how bothered the patient is by the condition. Maybe it is, but it's still a good

question.

CO-CHAIR NERENZ: Okay, let's keep moving along, because I don't know that there's any way, other than the point being made, it's all on the table. I think we've got Sherrie, Sean, you were up. You still up?

MEMBER O'BRIEN: Yes, pretty much asking the same question.

CO-CHAIR NERENZ: Okay.

MEMBER O'BRIEN: Over time, are there people who are just more likely to interact with the healthcare system for reasons unrelated to the intensiveness of their symptoms.

And then those people are going to get flagged because they're going to show some type of secondary diagnosis claim. So is it going to be partly measuring how many visits a patient has for whatever reason?

CO-CHAIR NERENZ: Yeah, good point. Okay, Sherrie, and then Dave.

MEMBER KAPLAN: The recently completed CEASAR Study, this big national study of prostate

cancer, localized prostate cancer and different treatments for it, surgery, radiation, and the act of surveillance, when using Epic, the Epic Measure, patient-reported outcome, you can see that where you start pretty much conditions where you end up.

So if you're a surgical patient with coming baseline erectile dysfunction and incontinence measures, you're going to end up about a whole standard deviation off of where you started from.

But it's all conditioned of your baseline. So I mean the patient-reported outcome business is relevant in this case. What's concerning to me is the non-adjustment and your no-findings of differences by rates.

Because it looks like for, based on our data from the California Initiative of Physician Medicine and the CEASAR data, African American patients, A, are less likely to get surgery and B, if they do get surgery, they have crummier outcomes.

And so I'm a little bit concerned that you're not finding a race variable there, especially since African Americans patients tend, for other conditions, to go to lower quality -- can you -- have we -- what did you find with respect to race and how is that playing out?

MEASURE DEVELOPER: We did test for it and it did not come out significant.

MEMBER KAPLAN: Interesting. Even so, within surgical differences across that variable or independent of the other adjustments, you didn't find any differences?

MEASURE DEVELOPER: Correct.

CO-CHAIR NERENZ: Well, although it could be --

MEMBER KAPLAN: Interesting.

CO-CHAIR NERENZ: Although it could be one of the data sourced problems and the proper -- that's one, and getting care for it is another thing.

And if it's a disconnect between the two then you're going to show up in one data set

and not in another data set because the effect will show up in one and not show up in another, and it just --

MEMBER GLANCE: Just have a really, really quick point of information. In our other sessions where we look at the impact of SES, of race on a risk adjusted outcome rate, typically haven't seen a really big difference.

MEMBER KAPLAN: Yes.

MEMBER GLANCE: We've seen a really high ICD --

MEMBER KAPLAN: It tends to move the whole distribution and it doesn't show up in people in total to the distribution. I get that.

MEMBER GLANCE: Yes.

MEMBER KAPLAN: On the other hand, it's kind of surprising with the variable itself because those same analyses, when you do, when you look at the individual variables, they're significant.

MEMBER GLANCE: Well, the variables, they're significant but overall the hospital

level --

MEMBER KAPLAN: Right.

MEMBER GLANCE: -- stuff doesn't really change pretty much.

MEMBER KAPLAN: Right.

CO-CHAIR NERENZ: Okay, Dave.

CO-CHAIR CELLA: So, I'm also concerned about the PRO being the sort of a standard that we all want and this being a surrogate for that or a proxy.

But I'm looking at the other side of the coin, the other side being, how many people were excluded because there were no claims? Like proportional to the whole -- it started with a population and then you excluded what percentage of people because, there were no claims or adverse events, these two adverse events.

Do you happen to know?

MEASURE DEVELOPER: No, we do not look at that.

CO-CHAIR CELLA: I don't, I'm guessing it's a high number. And, it's probably a lot of

people that have erectile dysfunction and/or urinary incontinence and just aren't getting treated for it.

And then a related concern that, you know, if you're a urologist that really wants good numbers, you might say well, it's just normal to have that erectile dysfunction.

We can help you just learn to live with it and not encourage a visit. I'm worried about that side of the coin, that this could actually tip healthcare away from paying attention to it, because there's a claims-based punishment, a claims-based penalty.

So I'm not necessarily expecting an answer. I'd like to know that number because I'm guessing that number of non -- no claims patients might be pretty high.

CO-CHAIR NERENZ: Okay. Just to clarify and following on that point and I may have misunderstood. You used the word excluded, if I'm in the denominator of this and in the subsequent year --

CO-CHAIR CELLA: You have to have a claim.

CO-CHAIR NERENZ: Well, but you have a -- yes, but what kind of claim?

CO-CHAIR CELLA: An adverse event claim, right?

MEASURE DEVELOPER: Right.

CO-CHAIR CELLA: Yes, because they only look at people with claims.

CO-CHAIR NERENZ: Oh, thank you, okay. So, no, that's important. Okay. So, if I would have left no claim for complication, I had the test and then the subsequent claims --

MEASURE DEVELOPER: In the numerator.

MS. WILBON: Yes, the denominator is that they had prostate surgery, open or closed, and survived at least one year.

CO-CHAIR NERENZ: Okay, so --

CO-CHAIR CELLA: So the denominator claim doesn't have to be in the denominator?

MEASURE DEVELOPER: Not in the denominator, only the numerator.

CO-CHAIR NERENZ: Okay.

CO-CHAIR CELLA: What is that proportion? What is the numerator over the denominator?

MS. WILBON: What is the proportion that?

CO-CHAIR CELLA: What percentage of the people that are alive for a year after prostate surgery have claims?

MS. WILBON: Had any claims at all --

MEASURE DEVELOPER: I don't, we can get that information, but I don't have it at the time.

CO-CHAIR CELLA: But you see why I think it's important, because if it's low then it's a real underrepresentation of the problem because people just aren't getting treated for it.

MEMBER TEIGLAND: And, did you require a minimum pre-index period, pre-surgery period as well?

MEASURE DEVELOPER: Yes, one year.

MEMBER TEIGLAND: One year also?

MEASURE DEVELOPER: Yes. One year before, one year after.

CO-CHAIR NERENZ: Okay. We are close to time here. If anybody's got something crucial to a vote? Jack.

MEMBER NEEDLEMAN: Yes, just we've been having a lot of discussion about the clinical logic of the way this is measured.

And, if the need, if the measure gets to the standing committee, I would encourage either the standing committee or a tech with urologists to really review the validity of the clinical logic behind this measure this way, which I think goes, the clinical issues go beyond our capabilities in this room. Thank you.

CO-CHAIR NERENZ: Okay, no cards up. Let us call the question now. This is now Subgroup 3 voting for real, everybody else, Survey Monkey.

CO-CHAIR CELLA: I object to referring to the shadow votes as monkeys.

MS. OGUNBEMI: We are now voting on the reliability for Measure 3478. The voting is open. Your options are high, moderate, low and insufficient.

CO-CHAIR NERENZ: That's the first time I've done this where it says clear last response. Does that mean you want to correct it or it means you're ready to move on?

MS. OGUNBEMI: If you clear it, that means you take it away.

CO-CHAIR NERENZ: I just wait for you to move us along. Okay.

MS. OGUNBEMI: Yes.

CO-CHAIR NERENZ: Okay, thank you.

MS. OGUNBEMI: So, we have, we were supposed to be expecting five members voting and we have four votes, John Bott, Marybeth Farquhar, David Nerenz, Eugene Nuccio, and Ron Walters.

MS. WILBON: Ron, I think --

MS. OGUNBEMI: No, we've got Ron. Oh, so we only took the form? All right.

MS. WILBON: Ron's here.

MS. OGUNBEMI: He's recused from voting.

MEMBER WALTERS: Oh, yeah, yeah.

MS. WILBON: He's recused?

MS. OGUNBEMI: Yes. Okay, consensus not reached. We have zero votes high, zero percent; moderate, 50 percent, at two votes; low, two votes at 50 percent and zero votes insufficient.

So Measure 3478 is consensus not reached on reliability. And now we are voting on the validity of Measure 3478. Options are, high, moderate, low, and insufficient.

Voting is open. And we have zero percent high, zero votes; moderate, 25 percent, one vote; low, 75 percent, three votes and insufficient, zero votes, zero percent. Measure 3478 fails on validity.

MEASURE DEVELOPER: Thank you.

MS. WILBON: Can I just make one last plug? If -- We are planning to take a look at the shadow votes tomorrow, so if you missed any

votes today on the Survey Monkey or subgroups that you, that you were not a part of, if you could just make sure that you have all your votes submitted up to now for all the measures that we've reviewed, that will help us tomorrow. Because we'd like to do some kind of on the fly analysis with you guys to look at, to see how the shadow votes compared to the actual votes and see, have some discussion about how those votes compared and maybe, what some of those differences might have been. So, if you could kind of catch yourselves up if you weren't able to keep up today for some reason.

That way tomorrow for the measures that we review, we can kind of pop right in and start looking at those and do a quick turn-around. So, thanks for doing that.

MEMBER ROMANO: Do you have any way of knowing which ones we haven't submitted or there isn't any way to know that?

MS. WILBON: We would have to do a quick export. And, it also depends on if you put

your name in or not, right. So, if you didn't put your name, we won't know who submitted what.

So it may be a little bit difficult if some folks chose not to put their name, so.

CO-CHAIR NUNEZ: So do we go to public comment?

MS. OGUNBEMI: So if anyone is on the line and would like to provide a public comment, now is your opportunity, we'll give you a couple seconds.

If your line is muted, you can press Star 7. We also have a chat function available on the web platform if you'd like to submit a comment that way.

(Off microphone comments.)

CO-CHAIR CELLA: Okay. We're going to wrap up. Aren't we?

MS. JOHNSON: Yes. So kudos to our fearless co-chairs who got us out of here on time.

CO-CHAIR CELLA: That was you, you cooperated wonderfully.

MS. JOHNSON: That was great. So pretty much, I haven't counted up, but we got through all the measures that we wanted to. I hope that the process wasn't too onerous.

What we would ask you to do for tomorrow is, if you have identified specific methodological issues that you'd like to talk about a little bit more tomorrow, we've been writing some down, but if there are others, you know, let us know or write them down yourself.

We will hopefully have a chance to talk about those tomorrow. We are also planning, I don't know if you heard Ashlie telling Dave and Dave that we're going to try to compile our shadow votes and just kind of see how that works.

We're also going to talk about, you know, how the process itself is working. It felt a little new and different, even than our separate pulse did, so we'll talk about that.

And then, probably more important than anything else, dinner. We do have reservations at Siroc, which is a, I think it's an Italian

place, just right around the corner.

I'll give you the address, because I cannot be relied upon to tell anybody how to get anywhere. It is 915 15th Street, so it literally is right up the street.

Our reservations are for 5:30, so that gives you time to go to your hotel, do whatever you want. So, if any of you would like to join us for dinner, we'd love to have you.

(Off microphone comments.)

MS. JOHNSON: Thank you guys, we'll see you tomorrow morning.

MS. WILBON: Thank you guys. Have a good night.

(Whereupon, the above-entitled matter went off the record at 4:54 p.m.)

T

C E R T I F I C A T E

This is to certify that the foregoing transcript

In the matter of: Scientific Methods Panel

Before: NQF

Date: 10-28-19

Place: Washington, DC

was duly recorded and accurately transcribed under my direction; further, that said transcript is a true and accurate record of the proceedings.

Neal R Gross

Court Reporter