NATIONAL QUALITY FORUM

\+ + + + +

PERSON AND FAMILY CENTERED CARE PHASE 3

\+ + + + +

TUESDAY
JUNE 7, 2016

\+ + + + +

The Committee met at the National Quality Forum, 9th Floor Conference Room, 1030 15th Street, N.W., Washington, D.C., at 8:30 a.m., Lee Partridge and Christopher Stille, Co-Chairs, presiding.

PRESENT:
LEE PARTRIDGE, Co-Chair
CHRISTOPHER STILLE, MD, MPH, FAAP, Co-Chair
BETH AVERBECK, MD, HealthPartners, Inc.
KATHERINE BEVANS, PhD, Children's Hospital of Philadelphia
SAMUEL BIERNER, MD, University of Nebraska Medical Center
REBECCA BRADLEY, LCSW, HealthSouth Corporation
JENNIFER BRIGHT, MPA, Momentum Health Strategies
DAVID CELLA, Northwestern University Feinberg School of Medicine*
SHARON CROSS, LISW, Ohio State University Wexner Medical Center
DAWN DOWDING, PhD, RN, Visiting Nurse Service of New York*
NICOLE FRIEDMAN, Kaiser Permanente
STEPHEN HOY, Patient Family Centered Care Partners
SHERRIE KAPLAN, PhD, MPH, University of California Irvine School of Medicine
LINDA MELILLO, MS, MA, CPHRM, CPXP, Spaulding Rehabilitation Network, Partners Healthcare System

LISA MORRISE, MA, Patient & Family Engagement Affinity Group, National Partnership for Patients
LENARD PARISI, RN, MA, CPHQ, FNAHQ, Metropolitan Jewish Health System
LISA GALE SUTER, MD, Yale School of Medicine, Yale Center for Outcomes Research & Evaluation (CORE)
PETER THOMAS, JD, Powers, Pyles, Sutter & Verville, P.C.
CARIN VAN ZYL, MD, FACEP, Keck School of Medicine of University of Southern California

NQF STAFF:

HELEN BURSTIN, MD, MPH, Chief Scientific Officer
KAREN JOHNSON, MS, Senior Director
ELISA MUNTHALI, MPH, Vice President, Quality Measurement
DESMIRRA QUINNONEZ, Project Analyst
KIRSTEN REED, Project Manager
SARAH SAMPSEL, MPH, Senior Director
SUZANNE THEBERGE, MPH, Senior Project Manager

ALSO PRESENT:

JACK FOWLER, PhD, Informed Medical Decisions Foundation
PAULETTE NIEWCZYK, MPH, PhD, UDSMR*
KAREN SEPUCHA, PhD, Health Decision Sciences Center, Massachusetts General Hospital
MATTHEW WYNIA, MD, MPH, FACP, University of Colorado, Center for Bioethics and Humanities

* present by teleconference

TABLE OF CONTENTS

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

8:40 a.m.

CO-CHAIR PARTRIDGE: Good morning, everyone. This is Lee Partridge and Chris Stille and the NQF staff and the members of the PFCC Committee. We're all bright-eyed and bushy-tailed, ready to go to work.

I did want to mention just one little housekeeping thing this morning. I think I probably misinformed a couple of you yesterday who were talking of taking the subway to the Reagan Airport this afternoon. The Red Line, which is the line that's right around the corner from the Mayflower, the station right around the corner from the Mayflower Hotel, is single-tracking from 10 to 3 or something today, or maybe all day. Anyway, it's a mess.

And the simple and most direct thing to do is to walk out of this building, walk out of the elevator, go straight ahead, turn right on 15th Street, walk down two blocks and you will come to a Blue Line Metro stop, which is called

McPherson Square. And you get on a Blue train in the direction of National Airport. And make sure you read the pylons so that you don't go the wrong way. You'll end up out in Maryland. And it's a straight shot and very easy. But please stay away from the Red Line.

Suzanne, I think, wants to just tell us one bit about the slides that you've got.

MS. THEBERGE: We just have a quick summary of what was decided yesterday. We thought we'd go over that briefly this morning before we got started. The Committee recommended Maintenance Measure No. 0420. And then the Committee also recommended the three CoreQ measures. Did not recommend the home and community-based services experience of care measure. Recommended three of the functional change measures, and did not reach consensus on two and did not recommend one. Those last three were the LTAC functional change measures.

And that's the summary of the discussion from yesterday. We had a great discussion, lots of interesting things came up. And today we've got two measures on shared decisionmaking.

CO-CHAIR PARTRIDGE: Right. Then we will proceed to consider Measure No. 2962: Shared Decisionmaking. The developer is Healthwise.

Are they here with us? So if you'd like to come up and take a seat at the table. Good morning.

DR. FOWLER: Hi, I'm Jack Fowler and I'm with the Informed Medical Decisions Foundation, which is now part of Healthwise. And I'm here to propose a measure of the process of patient and provider interaction and the extent to which it meets standards for shared decisionmaking.

Just to give you a quick background, because it felt like a long time, the foundation started trying to get patients informed and involved in decisionmaking in 1989. And the way we started doing that initially was to create decision aids so we could get patients to the point where they understood as much as their providers about the things that mattered in making a decision. We focused on preference-sensitive decisions; that is, where there was more than one reasonable option to deal with whatever problem they were dealing with.

And so we spent 15 years working on how to do that and doing it as well as we could. And at some point our board said, well, you've got to show that you're making the world better. We had tested individual things. There have actually been lots and lots of randomized trials, as you probably know, of good decision aids and how they impact people's knowledge, et cetera. But they wanted to know how we were doing more broadly. We wanted to change the world. And so we set out to see if we could document how decisions were being made and in a variety of places.

So, we hooked up with a group at the University of Michigan's Survey Research Center. And also there are some talented medical decisionmaking folks there: Peter Ubel and Angie Fagerlin and Brian Zikmund-Fisher. And we did, I think, what was the first national survey of how decisions were made.

And so they did actually a random digit dial sample of adults 40 and older, and we asked them about 10 decisions that happen often enough that actually a sample of 2,500 would have a reasonable sample of them. So we had three cancer screening things that happen a bunch. That's mammography, PSA, colon screening. We had three meds. And then we had actually four surgical interventions on cataracts, hip/knee replacement and back surgery.

And one of the things we had to do to do that survey was to develop what do we want to know about a decision so we can characterize how the decision was made? And after talking to a lot of people, we ended up with three things that were sort of essential.

One is that patients need to be informed. So we had to measure whether or not they understood the issues. Second is they had to have a meaningful interaction between provider and the patient so that the patient's voice had a chance of being heard in the decisionmaking process. And the last one is that ideally the decision itself would align with the goals, concerns, and priorities of the patient themselves when we got done with this process.

And so we worked -- this was 2007 and '08 -- to develop measures of all those things. And we did them first in the decision study. And we and others who worked with the foundation have been building on those measures ever since. And they are, I think, three different constructs and they each have to have their own measures and they each have a role to play.

So I'm going to talk about the process one. And we've been using that in several different settings. We have repeated the decision survey we did with Michigan and we've used it in a variety of places where we were studying how decisions were made. And one place in particular that also gave us wisdom for what we're doing -- we hope it's wisdom -- we were also looking for sites in which we could see what it took to make shared decisionmaking a routine part of clinical care.

And we asked for volunteers of clinical sites around the country who'd be willing to make a special effort to try to make shared decisionmaking routine. And we picked -- we actually ended up working with about 10 places around the country, in all kinds of settings, who would use decision aids when it was appropriate and who were committed to shared decisionmaking. And they became a kind of learning group for us.

And we collected a lot of data. One of the deals they had to make, we provided some support and they had to collect data as part of their commitment to us. And so we have a fair amount of data from them on all those topics: knowledge, alignment of decision, but also on the decisionmaking process. And so that's the background on which this measure came about.

We did not set out to build a measure, to create a measure of quality of clinical care per se. We had been working on how to characterize the kinds of decisions that patients make and the way they're made and whether they're made the way we'd like them to be made.

At some point we felt like we had enough data to say there's not another measure out there, that we know of, that if someone wanted to say, is this clinical site doing a good job of decisionmaking with their patients, I don't think there's another measure. And we decided we had enough data to make the case that this would be a reasonable approach to measuring that, if that's what you wanted to do. So that's why I'm here today.

CO-CHAIR PARTRIDGE: Our lead discussants are Katherine Bevans, Jennifer Bright, and my colleague Chris. Which one of you wants to lead off? Jennifer?

MEMBER BRIGHT: Okay, then.

(Laughter.)

MEMBER BRIGHT: Just because I haven't had my first cup of coffee yet. That's why I get the top.

So, this is a new measure, and as Jack said, designed principally to demonstrate the value of a shared decisionmaking approach. So it's a process-focused measure. We focus on assessing the after-the-fact, after-care delivery, after the provision of a surgical intervention for one of seven surgeries, to assess what level of shared decisionmaking occurred.

Generally speaking, I was just reading back through the comments to make sure I captured everything, but I think Healthwise provides evidence using the Dartmouth Atlas and some other resources to provide evidence of value, why patients and providers would find this measure of value to demonstrate what shared decisionmaking looks like, which I think I -- I know I feel is certainly an important goal of developing a measure.

I think, just as an overarching commentary, some of the questions have to do with how such an instrument is implemented. So, issues around time lag and feasibility of implementation. I think there were some questions about how missing data is handled. And I know I had some questions about -- there were a lot of references to the impact of decision aids and how they boost a response of shared decisionmaking, and yet your instrument -- and I know we're not supposed to talk about the instrument necessarily -- but the instrument doesn't actually ask about decision aids, and neither does it ask about whether a patient is asked about their goals or preferences for treatment.

And so I'm curious as to why those things were omitted, when your evidence that you're presenting as rationale for doing this measure talks about how both of those things significantly boost a patient's sense of being involved in shared decisionmaking. So those are the top line ones.

MEMBER BEVANS: So, just to add a few things. I would like to revisit and kind of echo what Jennifer said, in part because it was a comment that was a comment that was I think made in public comment about the omission of direct assessment of people's treatment goals and the degree to which the treatment actually addresses those goals. So, just to come back to that maybe in our discussion of importance.

But I just wanted to make sure that I understand fully that the measure itself is comprised of four items, I believe, and there's one that's sort of interchangeable based on the condition.

DR. FOWLER: Yeah, it is four items. It depends a little bit on the decision as to sort of how many options there are, is kind of the best way to phrase it. So we have two ways of asking the question about options presented.

MEMBER BEVANS: Okay. And then each item is scored zero to one, in some cases dichotomizing a Likert scale. And then the measure, it's a sum of those. I'm just kind of confirming so we're on the same page. But the measure is the sum of those four item scores and then be aggregated by calculating the average at the provider level. Right?

DR. FOWLER: Yeah, they're two yes/no items which are one/zero items. And there are 2/4 response alternative items that we dichotomize, as you say, and you get a point if you gave one of the top two answers and you don't get a point if you gave the bottom two answers.

MEMBER BEVANS: Okay. So lastly, again, before we get into the specific criteria, the information that you have to present to the Committee today is based on studies done with 266 patients, is that correct, from four sites with multiple treatment decisions or conditions in each of those sites.

DR. FOWLER: I noticed that comment about, are we leaving out multiple decisions, and you've got to ask these questions one decision at a time.

MEMBER BEVANS: Yeah, I just wanted to confirm --

(Simultaneous speaking.)

DR. FOWLER: And mostly these are, you know, if you're thinking about having your knee replaced, you may be doing some other stuff, but that's probably the biggest thing in your mind.

MEMBER BEVANS: I just wanted to confirm the data that are presented that are those 266 patients, is that right?

DR. FOWLER: They vary by the topic.

MEMBER BEVANS: Right. Total, though?

DR. FOWLER: Right.

MEMBER BEVANS: Okay.

CO-CHAIR STILLE: Great. Thank you. And I'm excited to see a measure on shared decisionmaking, because I think, from a policy standpoint, there's really very, very little out there and I think anything that we can do to move things along is good.

I had a couple of questions. The

first is that the level of measurement, the level of reporting, is this meant to be kind done at the provider level, or what's your sort of intent with this?

DR. FOWLER: Yeah, I mean, what we're doing is -- obviously, most of the work we did was at the individual level, and that's where we did thousands of interviews and all that stuff. But I think we have enough data to show that you can do it at a clinical site level. And you could do it at a clinician level, if you had the data.

And basically what we did here was pull together what we have from different sources to say, can we make the case that if you do it as a way of assessing the clinical care patients got, that you can get a meaningful answer?

And we thought that -- one of the challenges, as you've probably all wrestled with, when you're to validate something that you can actually say that we're measuring better clinical care, whatever that means, you've got to have a theory about, well, how would you know that they were doing something clinically that ought to show up?

And we thought that these sites that we're working with that are committed to shared decisionmaking were -- well, it was reasonable to say these people were trying to do shared decisionmaking more than the average person. And so saying that this measure showed up differently and more highly in those sites than they did in other places of usual care seemed like a reasonable body of evidence that this was a valid measure of clinicians doing the right -- doing this kind of thing better.

CO-CHAIR STILLE: Okay. I kind of wanted to know where to look for variability, whether it was within providers or among, between providers, for example.

Okay. Right. It's seven common procedures. Being a pediatrician I always say, all right, we can use more pediatric measures, but obviously the limitations are what they are with those procedures.

This is kind of getting into validity, and maybe I'll mostly save this for that, but they were all people that had the procedure. And so it excludes people that didn't have the procedure. And I wanted to kind of know just some information about why you did that. I think I can maybe guess, but I'd like to --

DR. FOWLER: Yeah, well, that's one of the things I wanted to talk about, was the selection of the procedures and the patients and why not.

CO-CHAIR STILLE: Right.

DR. FOWLER: And one of the biggest problems that we have found, and it comes from our work with the shared decisionmaking sites that we've been working with, is identifying people, reliably, who faced a decision and didn't do anything. It turns out to be virtually impossible. They do not document. I mean, you can go in and have a conversation about your knee, and if walk out and go to physical therapy, there's not necessarily anything that anyone could ever find that says you did -- we are most interested -- we kind of have a bias that thinks that the people who didn't get the intervention are the successes, in some cases.

CO-CHAIR STILLE: Sure.

DR. FOWLER: But we cannot figure out a way reliably to identify those people. And if you're going to have a measure of clinical care and how a particular site does it, you have to have a reliable way of getting the same population of patients who've had the same experience.

So the only thing I know that I can do reliably is, if they cut them, it will be in the record and then you can survey them. And I'm willing to make the case that those are the people you -- ethically, you want to particularly make sure -- you'd be particularly interested in what kind of decisionmaking process people who actually have intervention had. I think there's a case to say that that is a particularly group in that respect, though I would love to -- and maybe the future will bring a better way of recording things so we can do a better job.

The other thing, of course, we are always looking for is a lot of the interventions I would like to include here, sometimes it's hard clinically to make sure you can identify the people who actually have a choice. So we tried to get the ones where we thought most of these really did have an alternative that a reasonable person would follow. Whereas, some others, some clinical things can complicate it and it's trickier.

And, again, one of my goals was to say, can we identify a set of people that, one, we are confident should have had a choice? So, a rational person could have gone more than one way. And, two, that we can reliably identify across different sites so we're comparing the same kinds of patients from clinical site to clinical site, which seems is sort of essential if you're going to use it as a measure of quality of care.

So that's sort of how we got here. It's not perfect, but it seemed like a pretty good start.

CO-CHAIR STILLE: Great. Thanks. Those are kind of my major points. Thanks.

MEMBER FRIEDMAN: I have a question about potential people excluded. So would this include patients that can't speak for themselves when it comes to shared decisionmaking?

DR. FOWLER: Yes, it would exclude them. It just seems like the dynamics are just going to be quite different for those folks. And I would hope that there's a process for them, but it just seems like (audio interference) --

CO-CHAIR PARTRIDGE: Peter?

MEMBER THOMAS: Forgive me for being late this morning, but I read the material and I've listened to the conversation I was here for.

I totally understand the importance of -- that the measure and what it's kind of looking at. What I'm wondering if you could elaborate on is the nexus between being involved in the decisionmaking process and actually improving care, improving quality. So that's the -- I see that you've got in the evidence section a study of 2,800 patients and the like, but could you be a little bit more specific about the link between the two?

DR. FOWLER: Yeah, I mean, that case has been, I think, building in a lot of different circles, including NQF. I think if you go on their website they say shared decisionmaking is a good thing. But I the notion -- that we started thinking shared decisionmaking was a good thing, as someone mentioned, from the Dartmouth Atlas data years ago, where you see these wide variations from area to area that have got to be physician-driven. They're not patient need or patient preference. We have pretty good data on that.

So the argument that a patient, an informed patient involved in the decision should be a part of the decisionmaking, I think, comes part out of that.

The evidence about whether it's -- actually Karen Sepucha is going to be your next guest and has worked quite a bit on showing that shared decisionmaking makes decisions better. And that really has to do with how can you tell a better decision when you see one, which is not absolutely easy. I can say it really easily. You want the one that aligns with the patient's goals, preferences, and priorities. And that seems like a really great idea until you start writing the equation. And we've collected lots of data on what people care about and what they're concerned about.

And then it turns out that some people will make the decision based on one thing. It's not that they totaled up the pluses and minuses to do that. So we've had some difficult -- and we've asked people to summarize. That's another way to find out did you get a good decision. Would you do the same thing again? How would you rate the decision? And you do get some relationships between involving patients and how they feel about it. But, again, if they're not informed, for example, they can be pretty happy with a decision that probably doesn't make much sense.

So I think there's an ongoing need for more data about the relationship between the interaction between provider and patient and the net result, whether you get better health out of that, whether you get happier patients, whether you get better aligned decisions. I think that's still got to be a work in progress. And we haven't solved that problem yet.

I do know that if people aren't informed and they're not involved, then the patient voice won't be heard, and that doesn't seem like a good thing. That's the best I can do.

CO-CHAIR STILLE: Sorry. Just two quick technical things: the people on the phone are having a hard time understanding, so please speak directly into your mic. If you can hear yourself through the loudspeaker, chances are this is good.

The other thing is a couple people have their cell phones close to the mics and there's some interference. So just have your cell phone a couple feet away. Thanks.

CO-CHAIR PARTRIDGE: Ready to move onto evidence? Sorry. Lisa?

MEMBER SUTER: So I just wanted to comment. As a measure developer, this and the next measure are actually measures where I want to disclose that I'm not working on a measure related to this, but I am working on a measure of informed consent for CMS, which is tangential to what we're talking about. It in no way overlaps with what -- it's not a patient-reported outcome measure, but I wanted to disclose that information and defer to the chairs on whether or not I should vote.

But the other thing I wanted to notify you is that, in our conversations with patients about the informed consent process, if they are so eager to have us ask information about this, it's incredible. They are pushing for this kind of measure. They are pushing for us to ask. It's really -- you can't speak to patients about this process without them wanting us to go further than what we're doing.

So I just wanted to comment on the value of asking these questions and defer to you guys on whether or not I should vote. I can give you more specifics about the measure.

CO-CHAIR PARTRIDGE: The comments to my left were you're fine.

MEMBER KAPLAN: Yeah, I have a similar disclosure. I've known Dr. Fowler for decades, an undisclosed number of decades.

(Laughter.)

MEMBER KAPLAN: And I also have a -- I've developed for the medical outcomes study a global measure of participatory decisionmaking style. I was not involved in either of these two measures and haven't, sadly, working with Dr. Fowler for now probably more than a decade, either, but I just wanted to make sure everybody knew that. And I'm happy to recuse myself if there are issues of conflict.

CO-CHAIR PARTRIDGE: All right. Onto the first threshold question: evidence. And any comments on that topic before we proceed to a vote?

All right. Desi?

MS. QUINNONEZ: We are not voting on Measure 2962, Shared Decisionmaking. Voting is now open for evidence. Rationale supports the relationship of the health outcome, or PRO, to at least one healthcare structure, process, intervention of service. Option No. 1 is yes, option No. 2 is no.

We're looking for one more vote.

MS. THEBERGE: David, we need your vote. Can you submit it via the chat? Thank you.

MS. QUINNONEZ: All votes are in and voting is now closed. The results read 100 percent yes, 0 percent no.

We're now voting for performance gaps. Well, wait.

CO-CHAIR PARTRIDGE: The next item under evidence is that one: gaps, et cetera. Are there comments? Chris?

CO-CHAIR STILLE: There does seem to be a good gap and a reasonable score distribution. It's interesting to point out that the hospitals that you've been working with, since they're interested in this, may even be the best of the best. So I would think if there is any bias from having relatively small numbers, it would be in the direction where you'd actually see more variability among other hospitals. So even though the numbers are small, I felt that they were pretty good.

MEMBER DOWDING: Hi, it's Dawn on the phone. Can I just ask -- I just have a quick question, if that's okay?

CO-CHAIR PARTRIDGE: Go right ahead.

MEMBER DOWDING: It's just really to ask about the issue to do with the health literacy of the patients who've used the tool. I mean, I've looked at the samples and there's very, very small numbers of people who are in ethnic minorities. And I just wondered if there have been any account made for the health literacy of the patients are in the tool? Because there is some research to suggest that health literacy is actually related to an ability to be involved in decisionmaking. I mean, it's just a query really on whether or not the developers have got any ideas about how to deal with that.

DR. FOWLER: Actually, that's a topic we worried about a lot. And you're right that -- I mean, the data we have, particularly in the sites, are not as heterogeneous as you might want. It just turns out that's the patients they treat. I can't do anything about that. But we really have looked very hard, for example, at -- I mean, the hypotheses that people would have would be that these were people with less formal education, people who are older. That's another group of people like me. How would folks like that want to participate, and ethnic minorities?

And we have found very, very little evidence that that's true. And there's at least one published paper where we tested -- we didn't have ethnic minorities in that, but we had pretty good samples of low education people and older people, both about how important they thought -- one, how much they learned from decision aids; and two, how important they thought it was to be involved and informed in decisionmaking. And there absolutely was no difference.

In fact, older people, they benefit a little bit more, because less educated people started out with less information. So they benefit more from efforts to make the interactions informative. They actually raised their level of knowledge a little bit. And they respond really positively.

There may be some groups around, I mean, we've been hearing that don't want to share in decisionmaking and want to leave it to the doctors, but we just keep finding that that's a stereotype. And I'm not saying that if you don't have -- you know, one of the things that's really important is to have sources of information that are well-structured for people and don't make assumptions about their literacy levels and their prior information. And that's a really important part of getting information in the right form.

But as far as participating, we just have not been able to find evidence that there is resistance to participating in any particular group.

CO-CHAIR PARTRIDGE: Katherine?

MEMBER BEVANS: That's very interesting, first of all, your comments about that. So, thank you.

In thinking about gap, one point that you note in your report is that there is -- it's not quite a ceiling effect -- but a higher mean for a particular group; that is, those who have undergone surgery or are considering surgery for lower back pain. Which sort of makes sense, and you justify that by saying there may be less of a choice in that situation.

DR. FOWLER: Actually, I don't think so. I was surprised a little bit, but back is actually a surgery -- when you talk to orthopedic surgeons, in particular, they're worried about back surgery. And the majority of people who have back pain I think are not amenable and don't benefit from surgery. And so I think they work pretty hard to make people know the pros and cons. At least, most of them do. I'm pretty convinced that that's kind of real.

And so I didn't -- you know, and of the 10 procedures that we did -- or 10 decisions that we did in the decision study, the backs were by far the best. And so I think that's probably real, as best I can tell.

MEMBER BEVANS: Okay. Yeah, so I guess the more general question for me is, as you pointed out, the questions are really most pertinent to people in situations where there is a surgical intervention for which there is an option, right? You have either the option to, yes or no, you know, complete the intervention, or there are multiple choices.

So I guess I'm wondering for this measure whether that is a component of your specifications, that it may be most appropriate, or really only appropriate under certain circumstances, or whether you with the measure intend to provide any guidance around that point, sort of who it's appropriate for, who it may not be.

DR. FOWLER: Yeah, I mean, I want to make a distinction between who I think it's -- I think shared decisionmaking is appropriate for just about everything. I mean, I'm a bit of devotee about that. We picked these because, some combination of we thought that we could reliably sample people who had made a decision, and therefore for whom the shared decision model was definitely appropriate, and for which we had data.

I mean, for example, I would love to

figure out a way to do this for mammograms, for PSA testing, for taking statins. I mean, there are lots -- of taking depression meds, having cataract surgery. There are lots and lots of decisions that I'm pretty sure are overused and where having a standard of shared decisionmaking should be routine. I just either didn't have the data or was worried about being able to identify the decisionmaking moment.

Pills, for example, are really a hard one because it's sort of an ongoing thing. And did you have a decisionmaking visit and did it meet standards this last time? But you've got to sample it. Trying to sample across some period of time gets really confusing about answering questions.

So I hope this keeps growing and we find more and better ways to use these questions, but this seemed like a start.

CO-CHAIR PARTRIDGE: Stephen?

MEMBER HOY: Yeah, forgive me, I don't know where this comment falls in the process, but

I'm going to throw it in here. And I mirror Chris' excitement to see a shared decisionmaking measure come up. But my biggest concern -- I'm looking for you to ease my concern -- is that the priorities and goals of the patient influence the shared decisionmaking, and that, to me, good shared decisionmaking is when the provider and the patient have access to the same information.

So that means that my goals and priorities are understood by my clinician who helps me make that decision. And patient education, knowledge, information around their decision is an important piece of that, but I would like to see somewhere where the priorities are engaged by the clinician. And that may be at another level, but, again, I'm just looking for some information around that.

DR. FOWLER: Okay. Great. I absolutely couldn't agree with you more that having the patient's goals and concerns and priorities and preferences on the table in part of the decisionmaking process is the best. The best decisionmaking I know about was done at Dartmouth in the breast cancer center where the woman would see a decision aid and then she would actually fill out a questionnaire about her goals, concerns, and priorities that would go to the physician. And that would be part of the conversation that they had. And when you when we get the round trips all set up with the medical records and we're collecting data like that on a routine basis, maybe we'll have the goals and concerns integrated routinely in that way. And I think that would be fantastic.

We didn't happen to have the goals and concerns question in the initial set of questions. We asked about the process. And I mused about that some. And I've got a lot of data that don't have that question in it, so that's why it's not there.

I kind of think that -- I think making sure you ask the patient what the patient wants and have a discussion of the pros and cons -- you would hope discussion of the pros and cons would be a time when the goals and concerns would get on the table as part of it. It's a crummy discussion if it's the doctor talking at the patient. Then it won't achieve that, and then all we've got is the patient expressing what they wanted if they did that. I think it's an essential part. We also -- parsimony was kind of on our mind, and short seemed like it was better.

So I would love it if you'd go out tomorrow, add that question, and collect some more data and see how much it adds and whether it would be characterized. I think it would be super. It absolutely fits the model and it could absolutely have been there. So I don't push back about it at all. You don't have it because that's not the way we collected the data, so I can't do that today. But it's key and it should be part of it.

CO-CHAIR PARTRIDGE: Okay. Katherine, do you still have a comment?

All right. Are we ready to vote on gap, et cetera? You think so. Desi?

MS. QUINNONEZ: We're now voting for performance gaps on Measure 2962, data demonstrated considerable variation or overall less than optimal performance across providers and/or population groups. Option No. 1 is high; option No. 2, moderate; option No. 3, low; and option No. 4, insufficient.

Looking for one more vote. Okay. All votes are in. Voting is now closed. The results read 53 percent voted high, 42 percent moderate, 5 percent voted low, and 0 percent for insufficient.

CO-CHAIR PARTRIDGE: Dr. Fowler, I think people like this measure.

(Laughter.)

CO-CHAIR PARTRIDGE: I have to say that I have had a little bit of -- I've had an opportunity to work a little bit with you and talk with your colleagues about shared decisionmaking, with particular reference to C-sections, and it's always a pleasure to work with you all.

And we are now going to move on to a discussion of reliability and validity, our old friends, and scientific acceptability. So comments? Katherine?

MEMBER BEVANS: First of all, I want to note that your work on reliability, in particular the observations, is really, I think, wonderful, and I wish we'd see more of that in measures that are presented here where there are external coders looking at tapes of actual interactions between patients and clinicians and verifying whether or not a person's report of that process is a reliable report. I think that really strengthens the case for its reliability.

As you note, of course, at the provider level the reliability coefficient is somewhat low. And you note that that could be related to sample size, but that's something that hopefully will be mitigated with some more data.

CO-CHAIR STILLE: And, yeah, I'd like to echo what Katherine said. I think if this were to go so far as to be benchmarking providers against one another, or institutions against one another, we'd probably need a lot more sort of large-scale data, but for the moment I think what you have is good.

CO-CHAIR PARTRIDGE: Sherrie?

MEMBER KAPLAN: Yeah, Jack, that was actually my question. If this is supposed to be used at the provider level and you only have six -- it looks like six clinical sites, even though they're all around the country -- if this is intended ultimately to be used at the provider level, you'd want to do an intra-class correlation coefficient looking at between versus within provider variability, just to make sure that there isn't a lot of noise internal to these sites being measured.

So is that the plan? Or how do we evaluate? Because one of the measures coming up is a clinical measure to be used at the site level. This is supposed to be considered at what level? So we're at the patient level or --

DR. FOWLER: Well, I mean, I think

whatever site -- I mean, I think any delivery unit that you've got data for -- I think you can do it for a clinic, which might be a bunch of providers who did a thing, or an ACO. Or if you had enough cases, you can do it at the provider level. I think it can be -- we only had the data we had, and so that's why our samples are small. And we had to have them doing, obviously, the same procedure.

So we actually had a lot more patients, but in order to get replicates -- enough patients and providers where we could make replicates, and then look at the correlations when you had samples both -- we had what we had, and it wasn't overwhelming, as Christine said. But I hope we're going to collect more. We're trying to use this every chance we get.

MEMBER KAPLAN: So, Sarah, this is another one of those ones where a phased approval process, you know, this is a new measure in development, this is a more progressive measure. And as you keep going, you'd expect more of it and more data to come out, but help us in the group sort of understand where -- this is clearly a new measure. It's something that is going to have and need more testing before it's used at different levels. But where are we with that issue?

DR. FOWLER: Yeah, I mean, we haven't been mainly using it to evaluate clinical sites, so we had to use data that we had that we collected for other kinds of purposes. So that's where we are at the moment.

MS. SAMPSEL: And NQF does not have a phased approach. So it's either -- and I would draw you back, though, towards previous discussions, previous decisions. We are at a point where we're facing new measures, and we're facing novel measures that we haven't seen before. And so I do think NQF is still grappling with how we're going to deal with that.

One of the things -- and we can talk about it a little bit later -- a totally separate project on shared decision aids that NQF that is funded for right now, but that doesn't help you here. So I guess what I would encourage you all is to think about kind of the results as they were, your past decisions on other measures with similar types of data. And these are also things that, you know, they go into the report and we'll constantly be watching.

But I would say, you know, when Jack comes back in a year with an annual update, and then in three years for maintenance review, the review process gets a whole lot tougher because we're going to want more of that data.

CO-CHAIR PARTRIDGE: Jennifer?

MEMBER BRIGHT: I'm going to betray how long it's been since I've been in grad school statistics class, but is there any importance to the number of responses that would be considered important to have a reliable score? You don't really speak to that, but since we're talking about sample sizes, like applications, measure, how many responses are really needed in order for it to be a meaningful representation of what's going on in the practice? And because of that, how can you compare practices?

Because let's not kid ourselves, this is important, at its core, for internal practice improvements, but we all know everybody will still look at it and say, how can you judge best approach, right? So is there an N that has to be met? Is that in your specifications?

DR. FOWLER: Well, I took a stab. As you know, the sample size thing is about what your standard errors are and then how confident you'd be that the differences that you observe matter, are big enough. And so, it looks like -- and it somewhere in the 50 pages there.

But if you look at the standard errors around these things, it's actually not that big. And if you had samples in the 50 to 100 range and you looked at how much difference we were observing between places that looked like they were doing it pretty well and usual care, that you would probably pick that up most of the time.

So I think that I feel comfortable

saying that, for what we're talking about, that getting 75 to 100 responses would be a good thing. And I actually -- one of my -- we are talking -- it's on my list of things that people asked about -- and maybe it's some other section, but I actually, of course, think non-response may be a bigger deal than sample size in this world these days.

I'm really worried about non-response everywhere and how the self-selection that shows up when your response rate is 25 percent -- who those 25 percent are and what's on their minds. It's different from the other 75 percent. And I think that's a big deal for almost any patient-reported measure.

So if I were going to spend my money on data collection, I'd take my 75 and then try to get as good a response rate as I could to generate that 75 as compared with going about getting a sample a little bigger. That's sort of where I am.

But I think the standard errors look

like 75 to 100 will detect clinically meaningful differences between places that aren't doing a very good job and places that are. And that seems like the key issue.

CO-CHAIR PARTRIDGE: If there are no further questions, are we ready to vote on reliability? Desi?

MS. QUINNONEZ: Voting is now open for the reliability of Measure 2962, reliability including precise specifications and testing. Option No. 1 is high; option No. 2, moderate; option No. 3, low; and option No. 4, insufficient.

Okay. Looking for one more vote.

All votes are in and voting is now closed. The results read 11 percent voted high, 74 percent voted moderate, 16 percent voted low, and 0 percent insufficient.

CO-CHAIR PARTRIDGE: Okay. Moving on to validity. Comments? Katherine?

MEMBER BEVANS: Since you brought it up, response rates. Could you tell us a little bit about how that played out in these data? Because I didn't -- maybe I missed it, but I didn't see it in the report. And I never know whether to bring up response rates as part of validity, as a threat to validity or feasibility, but here it is.

DR. FOWLER: Yeah, and there's a range. And the reason you didn't have them is because all these different data sets and response rates are different everywhere.

The two national surveys -- the trend survey, it was based on a panel. It was high, but they were already in a panel, so that doesn't -- I mean, like, they were 85 percent or something. But the key ones are the ones on the clinical sites. And they had a range from 25 to 60 percent. And we talked about one of the things that maybe you wrote in terms of, what are the issues about how do you collect data and what do you do?

And there's no doubt that, one, how clinical sites sort of present that they care about getting this back. If somebody actually handed something to somebody and said, "we read it" or "we're going to send you a questionnaire about things and it would really be helpful if you'd send it back," that really helps. And to follow up and say, "could you send it back?" helps, too. And some places really care about it and they do better and some places aren't.

The average CAHPS survey these days is 27 percent. I think almost all the data -- I think all the data that's represented -- there might have one 23 percent or something -- but I think all the day that's represented in here are in that ballpark. And the average is higher than that, but it's all over the place, to be fair. And I don't know. I can't tell you exactly how the answers would have been different if the response rates were higher. Again, sorry. But it's a good question.

CO-CHAIR PARTRIDGE: Jennifer?

MEMBER BRIGHT: I know this is problematic, too, but you yourself said that the response pool was pretty homogenous, old white men. And except for the breast cancer, right? Good point.

DR. FOWLER: Yeah, and the mastectomy, by the way --

(Laughter.)

MEMBER BRIGHT: But it does bring up the point -- I think one of the big questions that everybody has is, okay, shared decisionmaking is still a pretty novel concept, and how do you develop a process to do it well and then measure its impact? But we also know that there's a high level of disparities in decisionmaking, in communication, language barriers, all those things, not just health literacy.

So, you know, what did you examine about that? Is there anything that needs to be talked about in terms of how valid these results are given that it was such a homogenous respondent pool?

DR. FOWLER: Yeah, I mean, the sample

about being old white men, that was only 65 percent men. That's a lot of men, I agree. But it did have two of the procedures, too, as well as the sites. And we need to expand this kind of measure to the decisions that all kinds of people make. Clearly, actually, pregnancy-related decisions are ones where you get -- with a different population, for example.

MEMBER BRIGHT: Yes.

DR. FOWLER: And that's a place where people are working hard on having more informed and involved decisions about C-sections and induced deliveries and things like that. So, as I say, I mean -- and we have collected data, as I say, on other topics, but not in the form that I could present them to support the validity or where I thought we could reliably get people.

So there's nothing conceptually about why these questions shouldn't work. And we have collected the data, as I say, about mammography and statins and stuff where you get really different samples. So I don't think there's any reason to worry that they don't work to capture what people have to say.

And by the way, the coding studies where we had coded the interactions and compared them to the response answers, those were done with breast cancer. So that was a different population, too.

So I don't think there's any reason to think that these questions don't capture well what the interaction is between providers and patients of a variety of sorts. There may be -- I mean, I can't do -- I don't have the data, for example, about people with different languages and stuff like that, and that's -- I don't know where that is, but I think mainly I'd -- I think I'd have -- I feel pretty comfortable that they -- that there's not a basis for saying these questions only apply to -- or mainly work best with white guys who are making orthopedic decisions. I don't think that would be unfair.

And for example, PCI is one we're talking about, and those are among the worst decisions anywhere. I mean, those patients have no clue what's going on. And those are a lot of white guys. There's some women, too.

CO-CHAIR PARTRIDGE: Lisa?

MEMBER SUTER: I saw you did a lot of work in development with patients. Did you take this back to patients after completing it and ask them their opinions or have you presented the results to patients at the end of the process?

DR. FOWLER: Have we taken results from process scores and fed them back to patients and said what do you think about --

MEMBER SUTER: That and the questions that you finally came up with. Have you shown those questions to patients and asked them --

(Simultaneous speaking.)

DR. FOWLER: Well, we did do a lot of cognitive testing of questions. That's sort of a routine part of what we do. And at least what that means for us is asking people questions, having them answer the questions and then talking to them about how they understood the questions, talking to them about how they arrived at the answers that they had and whether or not -- and working with them about whether the answer reflected what it is they had to say about that. That's not saying whether -- that's not the same as question -- as whether they care whether the answer is, but at least we had a lot of patient input in terms of the design and the selection of particular questions that we did.

I don't have feedback -- and I don't know if it was your comment, but there was a comment, or maybe more than one, about -- the question was is this valuable to patients? And what I wrote was I think we got a lot of evidence that patients value being informed and involved. And that didn't say did patients think these are the right four questions? And I don't -- I did not ask -- I don't have data that -- or experience to where we asked patients whether they thought these were the right four questions. So, I guess the answer is no.

But I do think the issue about whether

that's the right role for patients to play is a little bit tricky because there's just so many different ways to go, the things you could ask people. And likely didn't ask them -- talk to them about their goals and concerns and that sort of thing, which might have sounded more like what they wanted to talk about. Anyway, the answer is I don't have experience with where patients voted on whether these were the right questions.

CO-CHAIR PARTRIDGE: Sherrie?

MEMBER KAPLAN: Thanks, Jack. To follow up on the content validity question, you didn't pull these out of the blue, though. These were -- you defined the domain of observables and you picked the -- just to follow up on Lisa's question, you saw a group of folks --

DR. FOWLER: Right, I think there's a pretty well-developed --

(Simultaneous speaking.)

MEMBER KAPLAN: -- validity estimation. Yes.

DR. FOWLER: -- structure of whether

the elements on shared decision ought to happen.

MEMBER KAPLAN: Right.

DR. FOWLER: And that's where these came from.

MEMBER KAPLAN: So those were --

(Simultaneous speaking.)

DR. FOWLER: Yes.

MEMBER KAPLAN: Okay. So my question is -- and this isn't meant to be a criticism. It's this is one of the few proposals, at least I reviewed, that actually looks at an effort to improve quality and the responsiveness of the measure to those efforts to improve quality. So in the shared decision making sites it looks like measures responsive to efforts to improve shared decision making.

So then my question is how did you randomize? Because if you randomize within site, it's a little bit different than if you randomize site. Did you do cluster -- how did you do that?

DR. FOWLER: Yes, I mean, I think interventions about changing clinical practice, it's hard not to contaminate. One of the reasons, I think people like historical controls kind of better so that you can sort of see how are things going first and then introduce something and then see if you change things. But it's the same guys and gals who are doing it.

And that takes a variable out, because once you start -- once a clinician -- the notion that a clinician is going to share decisions with patient A and then not B, or even that within a practice that half the doctors are going to be doing shared decision making and using decision aids and the other half are not, I think that really gets hard to do. And we've fooled around with that, but I think we've always felt like we had to randomize practices.

MEMBER KAPLAN: So you randomized practices?

DR. FOWLER: Yes.

MEMBER KAPLAN: Okay.

CO-CHAIR PARTRIDGE: Nicole?

MEMBER FRIEDMAN: My question is about

is this focused specifically on the relationship in shared decision making between the provider and the patient, or does it also include providers that work in care teams, and how -- yes.

DR. FOWLER: Yes, I think it includes anybody in. I'm not -- I think the future and getting this right is going to actually involve non-physicians in particular. And I think the wording is such that people -- and people do report if they have the conversations. It doesn't matter whether they had it with a health coach or a nurse or with a doctor. That all counts. And I think that's really an important part of shared decision making is expanding the team because it looks better.

MEMBER FRIEDMAN: And would there be an -- I'm sorry if this is more than use and usability. Is there training that would be developed and implemented to knowing if there's varying levels of comfort with having those conversations, especially if they go to health coaches or non-licensed --

DR. FOWLER: Yes, there are a couple training programs that have been developed now that -- both for providers of all ilks about how to do this. And there are online ones that are now available, a couple of them. I think AHRQ has one.

MEMBER FRIEDMAN: And did the people that were surveyed, did those providers have the training?

DR. FOWLER: No, they didn't, because these people actually sort of were the pioneers of sorts. They did have a -- we did have a weekly, a monthly kind of support group phone call so that at least a lead person from each site -- I don't know if -- people talked about the problems they had and how they solved them and successes and things like that. So we kind of made it a working seminar, but we didn't have the formal training then. But now there is training available and it's starting to be used.

CO-CHAIR PARTRIDGE: I am sorry.

Peter?

MEMBER THOMAS: I have a very brief question and it really doesn't honestly bear on these questions that we're answering here, but I'm just curious. Have you ever asked the pool of people that did identify in the affirmative of being involved in the decision making and answering those questions that you laid out positively -- have you ever asked would your decision have been different had you not felt that you were fully informed or had that --

DR. FOWLER: Boy, that's a hard question. I think the answer is no. Certainly we haven't done it in a structured way.

MEMBER THOMAS: You see what I'm getting at? I'm trying to figure out whether the process is going to --

(Simultaneous speaking.)

DR. FOWLER: Oh, we have anecdotes all over the place of people saying, oh, gosh. I mean, but we definitely have gotten feedback from patients who'd see decision aids, talk to their doctors and say, oh, my gosh, I didn't understand that or -- either because -- I mean, because they said, oh, gosh, if I -- now that I know that a herniated discs get better on their own in a year or two if I don't have surgery, why would I want to do that? Or now that I know PCI is not going to save my life and I don't have any symptoms, why would I want to have that?

I mean, we definitely -- there is the a-ha moments and it's separating the information from the conversation with the doctor, I can't quite do that.

MEMBER THOMAS: Right.

DR. FOWLER: But shared decision making, absolutely. All of -- the average of tons of -- not tons, dozens of randomized trials with patients who see decision aids and don't show that they're more conservative about interventions. That is, you get less -- patients opt for less intervention when they're informed in the U.S. and then under usual care. And that I think has been pretty well -- about 25 percent less fewer. Though probably depends on the mix of interventions that you're studying, but that's the average across let's say several dozen studies like that. So there's no doubt that being informed and involved has an impact on decisions on average.

And it tends to be in this country -- in the U.K. you get a few things that go up higher because their rates of some interventions are lower, but they're more conservative. But in the U.S. it's almost all due to have less intervention. No, not for everybody, obviously. It goes both ways.

CO-CHAIR PARTRIDGE: Okay. We ready to vote on validity?

Desi?

MS. QUINNONEZ: Voting is now open for the validity of Measure 2962 including specifications consistent with evidence, testing, threats addressed, exclusions, risk adjustment stratification, meaningful differences, comparability, multiple specifications and missing data.

Option No. 1 is high; option No. 2, moderate; option No. 3, low; and option no. 4, insufficient.

(Voting.)

MS. QUINNONEZ: Waiting for one vote.

(Pause.)

MS. QUINNONEZ: Can we have your submit your votes one more time just to make sure everyone -- all votes went through?

(Voting.)

MS. QUINNONEZ: All votes are in. Thank you. Voting is now closed. The results are 11 percent voted high; 79 percent voted moderate; 11 percent voted low; and 0 percent for insufficient.

CO-CHAIR PARTRIDGE: Okay. Moving on to feasibility. Comments? Sam?

MEMBER BIERNER: I want to ask, there was a question that was raised in the comments about what are the -- are you going to have any public reporting of this, or how is it -- other than being used at a hospital level or some other group level, is there any plans to have any public reporting of this information or use it more widespread?

DR. FOWLER: Yes, I mean, that's not a role we can play actually, so I don't have any particular plans. I think there are a lot of contexts in which people are thinking about -- they want to measure whether or not shared decision making is happening. And I think this is a reasonable way to do it. So that will vary by the context in which people want to use it, I guess is the right answer. I think it certainly could be publically reported, but that's not something that we can have a direct role in.

CO-CHAIR PARTRIDGE: Peter?

MEMBER THOMAS: So a similar comment to the one I made yesterday. When you're talking about mailing surveys and then following up with phone calls, that's kind of time-intensive and maybe -- I don't know expense. The data wasn't submitted for how much, how much expensive that is to do. But are there ways to take advantage of IT and other ways to gather this data, or is this just a different population?

DR. FOWLER: Yes, everybody wants to do that, and I think that's where we're going to be in the future. The response rate experience has been very -- quite variable there. I know of one place -- I think places that have some kind of savvy client populations that are kind of comfortable with their portals and do better -- and the ideal way to do this would be after a visit to have questions -- a letter go out and to get the patient to go to a portal or something, click on a URL and answer the questions there. And I think that will be the future probably for a lot of people.

But at the moment, as I say, the response rates have been much worse with that approach than they have been in -- in most settings though -- so if you were in a setting where you have a population that will work with that, I think that's fine. And we certainly collected these data online, so the questions worked just fine there. So it's really just a matter of feasibility and what a place can accomplish. But the one I don't think you should do is to have a nurse collect the questionnaires in the offices, because we actually have data that show that that really gets pretty distorted results.

But they're not huge samples that you're going to have -- that a place would have to do. And I wouldn't think you'd just do this continuously. In fact, I think you might even vary the topic from year to year. You might do hips one year and backs another year. So it might not be an astounding thing, but somebody -- if you do it by mail, somebody's got to do the work and have follow-up phone calls, and that does take some time from somebody, either from a practice or somebody they hire.

CO-CHAIR PARTRIDGE: Becky? Sorry.

MEMBER BRADLEY: Thank you. I guess I'm still struggling with the lack of information from patients who chose not to have surgery. And you've explained why those were not included here, but how does this relate to or overlap with CAHPS surveys and how are these questions different? Because if they were used in a CAHPS environment, then you would have data from people who chose one method over the other. So is there any discussion on harmonizing these?

And the other issue with that is we're hearing from patients that they're being surveyed repeatedly once they got home and they're kind of getting the surveys confused as to which environment they're being surveyed about again. So I was just wondering if that's been a discussion with your --

DR. FOWLER: Yes, well, obviously CAHPS is a survey that happens. And actually they have -- the patient-centered medical home instrument had some questions that were taken from -- were derived from this a bit. But the problem that they have in -- that you have in CAHPS is that the rates -- given the sample sizes, the rates at which any particular decision is made are so low that you don't get adequate samples.

Because if you sample even 300 visits from something, the rates at which you'll find anybody who made any of these kinds of decisions, you might pick up -- I mean, they couldn't even do specifics. They were asking about medication decisions, but they had to lump all the medications together. And the problem with that is that if you're summarizing across several decisions, the data become mush. The people just -- we really know you've got to ask about one decision at a time to get variance.

And so, I think the problem is you just can't get adequate samples that way. I mean, that is the way we got samples of people who didn't make the decisions is to have a really big sample of 40-year-olds. And for that you get a lot of people who talked about knee surgery or talked about statins who didn't make the decisions. I could not agree more that I want to know how those people are treated, too, but I don't know how to make it feasible to compare practices.

And if you think the burden -- well, the burden will be whatever it is for targeting people who have particular -- these interventions, but if you tried to make the samples big enough to find even this very small number of decisions that will happen often enough to justify a -- to get a reliable result on decision making, it will just be huge. So it's a practicality problem at the moment.

CO-CHAIR PARTRIDGE: Further comments? If not, I think we're ready to vote on feasibility.

Desi?

MS. QUINNONEZ: Voting is now open for feasibility of Measure 2962, including data generated during care, electronic sources and data collection can be implemented. Voting options are 1, high; option No. 2, moderate; option No. 3, low; and option No. 4, insufficient.

(Voting.)

MS. QUINNONEZ: Okay. All votes are in and voting is now closed. The results are 0 percent voted high; 63 percent voted moderate; 37 percent voted low; and 0 percent voted for insufficient.

CO-CHAIR PARTRIDGE: Moving on to usability. Discussion? Sherrie?

MEMBER KAPLAN: Jack, how do you see this being used?

DR. FOWLER: Well, I think the people who are thinking about -- I don't -- I mean, one obvious place that I think may get -- generate a need is in the accountable care organization evaluations, that somebody is going to have to decide whether shared decision making is happening.

And I think in that context the notion of -- I think I could see picking a procedure of the year or something and having -- today we're going to take a look at how PCI decisions are being made in samples. And so, we're going to go make people -- get surveys back from 100 people who had PCIs and we're going to find out how the decision making process went. Ideally, you'd also get questions about whether they understood what was it was about, too. So that's one kind of thing.

And because I -- a lot of people are talking more and more that shared decision making use is your standard. I don't know all the places that are happening -- there's all -- but -- and what the potential ways of evaluating are going to be, but I know we need some questions to ask if that's going to be important. And I think these are the best that I know about at the moment. And I hope we'll keep learning more about them and I hope maybe we'll even make them better then --

MEMBER KAPLAN: So it's mostly for quality improvement, you're saying?

DR. FOWLER: Well, in this context I think it's quality improvement. It's also needed for research all the time when we're trying to figure out how to stimulate better interactions between patients and providers and more shared decision making. And I think this is an obvious measure for researchers to use on a -- and I think if NQF says that this is an okay way to do it, I think that would increase the odds that they'll use this approach, and that would probably be better. I think that's my --

CO-CHAIR PARTRIDGE: Sherrie, I'm just going to comment here. I live in a state, New York, which is enamored of the concept of value-based purchasing. This issue lends itself to that arena because it's entirely possible you'll see disparities, as Dr. Fowler has noted, and those of us who love Dartmouth Atlas know, you can find communities where the incidence of say hip surgery is much greater than it is in another.

And it will certainly be very appealing to somebody crafting a QI program in an ACO or in a clinic to say, oh, here's a measure that can begin to help us understand what's going on. And I frankly -- I have a feeling this may be quite popular.

MEMBER KAPLAN: You are not going to get a pushback from me on this issue. On another hand, I was just interested to see the vision of the measure developer here and whether this was ultimately going to land in the accountability zone. And if so, obviously we're going to need more data and much more information about how this works in that space rather than quality improvement. But if it right now is being considered as a quality improvement measure, it helps us, I think -- or it helps me at least to understand what the use and usability issue is.

CO-CHAIR PARTRIDGE: Are we ready to vote on usability?

MS. QUINNONEZ: Voting is now open for usability and use of Measure 2962, including accountability, transparency, improvement, progress demonstrated, benefits outweigh evidence of unintended negative consequences.

Option No. 1 is high; option No. 2, moderate; option No. 3, low; and option No. 4, insufficient information.

(Voting.)

MS. QUINNONEZ: All votes are in and voting is now closed. The results are 32 percent voted high; 58 percent voted moderate; 11 percent voted low; and 0 percent voted for insufficient information.

CO-CHAIR PARTRIDGE: Last question: Overall, do we recommend this measure for endorsement by NQF?

Desi?

MS. QUINNONEZ: Voting is now open for the overall suitability for recommendation for endorsement of Measure 2962. Option No. 1 is yes; option No. 2 is no.

(Voting.)

MS. QUINNONEZ: Looking for one more vote.

All votes are in. Voting is closed. The results are 100 percent voted yes; 0 percent voted for no.

CO-CHAIR PARTRIDGE: Dr. Fowler, this Committee doesn't do that very often.

DR. FOWLER: Well, I definitely want to thank the Committee, both for your thoughtful comments -- and I also say that I'm dazzled by the amount of material that you all have to read in order to do your jobs. And I think it's very outstanding and you should be well-thanked for your service.

CO-CHAIR STILLE: Well have a five-minute bio and coffee break.

(Whereupon, the above-entitled matter went off the record at 9:59 a.m. and resumed at 10:07 a.m.)

CO-CHAIR STILLE: Great. Welcome back. We're going to proceed with the discussion of measure 2958 which is Informed, Patient Centered Hip and Knee Replacement Surgery Measure from Mass General Hospital in Boston. We have a measure developer here to talk to us. The discussants, primary discussants, will be Sherrie

Kaplan, Stephen Hoy, and Rebecca Bradley. And we are ready to go. So take it away.

DR. SEPUCHA: So good morning. My name is Karen Sepucha and I run the Health Decisions Science Center at Mass General Hospital. So I'm here to present Measure 2958, which is informed, patient centered hip and knee replacement surgery. And I think this is very complementary to the discussion we just had around shared decision making measure.

The whole purpose of shared decision making and the goal of shared decision making is really to ensure that patients are well informed and they're getting treatments that match their preferences. And so what this measure does is look at that and provide a way to assess that for folks making elective surgery decisions.

These are high-stakes decisions. Hip and knee replacement is a very common decision. And although there are appropriateness criteria to determine whether or not someone is clinically eligible for it, just because someone is clinically eligible for one of these procedures doesn't mean they should have it. Someone really should only have one of these surgeries if they are well informed and have a clear preference for it.

So what the measure does is try to evaluate the quality of those decisions and it is based, the performance measure is derived from survey instrument, a decision quality survey instrument that we have developed and used in several different arenas.

There are six items to the survey, five knowledge questions, and then one item is a preference, elicits patient's preference. This is the kind of information that really can only be ascertained from patients themselves. It is not routinely documented in a chart or anywhere else, so we really do need to be asking patients about this in order to understand.

And the evidence in the submission does provide some evidence around gaps in the measure. The studies that we have done have found that there are maybe about a third to a half of the sites that we were surveying could actually meet the criteria, so patients who actually were both well informed and received care that was consistent with their goals.

And then, there was also evidence that suggested ways to improve that, so sites that had used patient decision aids, who were engaged in more shared decision making had higher rates of this.

So we've developed these decision quality instruments, the surveys upon which this measure is based for several different conditions, about 14 different topics. This is the first performance measure that we're proposing, so this is new. This is new for us and I think as some of the comments noted, there's limited data so far that we have, but we are interested in increasing the amount of data to better understand how this performs, as well as potentially extending it to many other different topics.

I do think it complements the measure that Jack just proposed around shared decision making, which is sort of a general process measure, so what was the conversation? And this really complements it to say well, okay, we could have a good conversation, but did patients actually understand the information? Did they have a clear preference for this procedure? And those are things that we really want to know before we move forward with claiming that they had a good quality decision.   So I think I'll stop there and I'm happy to take obviously questions.

CO-CHAIR STILLE: Great. Thanks. We'll open the discussion. Sherrie, Stephen, or Rebecca would like to start?

MEMBER KAPLAN: We have a problem in decision making.

CO-CHAIR STILLE: You're sharing the decision to talk at the same time.

MEMBER KAPLAN: Okay, so thank you for that presentation. I have a couple of questions because this is a new measure and I was looking closely at the questions themselves. And some of them I note like after knee replacement, how many months does it take most people to get back to doing their usual activities? And things like the baseline state would make that a little bit interesting to try and interpret from a patient's standpoint.

If 100 people have knee replacement, how many will have a serious complication within 3 months after surgery related to the surgery? So there are some questions I have about question content. But then I also was looking at sort of the -- well, I'll come back to risk adjustment and the Woloshin numeracy and risk literacy kind of issues in just a second.

So this is a new measure and so the data, as you aptly pointed out, are somewhat limited. The reliability data that you provide are on sort of the test/retest reliability, for example, and the time interval there was like -- I forget what it was, four weeks? Four weeks, yes. And so that looked good.

The practice level though, if this is meant to be a practice level variable used at the practice level, what you would want is between versus within practice variation, not a split half reliability. And that may be an issue of the guidance you got from NQF. But that wouldn't be what we would ultimately be looking for if this is intended to be used at the practice level.

So my first question is is that the ultimate intent for this? And if so, where in the process are you with respect to estimating how good this is for practice level discrimination?

DR. SEPUCHA: So I think there is some flexibility, but do you mean provider level versus practice level?

MEMBER KAPLAN: No, patient level versus practice level.

DR. SEPUCHA: Patient level versus practice level. And I also did get some comments back from my statistician as well that might have some additional information on what the test was that she ran and what we were looking at. And we're also sort of limited by the sample size in terms of what we could say about different practice level things.

So I think ideally we would want to have it at the practice level, as well as at an individual level. As Jack was mentioning, a lot of the shared decision making does kind of go on at a practice level whether it's do you have access to decision aids? Do you have nurse practitioner helping with education? So these could reasonably be assessed at a practice level to see whether or not they're doing things well.

MEMBER KAPLAN: I keep going, Chris?

CO-CHAIR STILLE: Yes, absolutely. Keep going.

MEMBER KAPLAN: Okay, so a couple more things. Discriminant validity in terms of patients whose providers had the decision aids versus folks who didn't looked very positive and in that sense it's like responsive to an effort to improve quality in this area. But then I also noticed that it looked like, and correct me if I'm wrong, it looked like you were comparing the provider's knowledge to the patient's knowledge for evidence of discriminant validity.

And my question would be did you not or could you not have compared people with high numeracy to low numeracy on whether or not they got fairly good scores or not so good scores? Why choose groups that were going to be pretty much that far apart on this issue?

DR. SEPUCHA: So we did set up a couple of different measures or tests for discriminant validity. So one was a group who got a decision aid and a group who didn't, so those were sort of patient level and compared those and we found some significant differences on the knowledge questions.

The other, just because a lot of places weren't using decision aids, so we set up -- you know, we had healthy volunteers, so people who didn't have the condition versus patients with the condition versus providers. And part of that, obviously we expected a pretty big difference between providers and the patients. But we were also doing that in a sense to look at the content validity of the items. So if providers were getting the questions wrong, then there's probably something wrong with the question in terms of -- or it was either controversial or -- you know, we also wanted to make sure that actually these questions had, quote, a pretty clear answer that people could consider correct.

MEMBER KAPLAN: How many of the doctors got it wrong?

DR. SEPUCHA: So, it actually depended -- we did a multi-disciplinary group of doctors. The surgeons actually, for this one, the surgeons and the specialists pretty much got them all right. Primary care doctors it was a little bit different though. And so that was interesting. And we did this across -- we've done this across many different topics. I think the biggest one was NDC. The primary care docs had a very -- their knowledge was not as good as we would like it to be.

MEMBER KAPLAN: Thank you. One final question. The other thing is this measure is administered two years, up to two years after the procedure? Is that accurate?

DR. SEPUCHA: So we've done it in different ways, depending on the study and the proposal here, I think, is similar to what Jack was wrestling with is figuring out how do we reliably find a sample of patients? And so we can reliably find a sample of patients who have had the surgery and then the studies that we have done have looked up to two years after and surveyed patients and so we have a good data set there.

In our own institution, we've actually been surveying patients the week after their visit with the surgeon, but we can find them a lot easier and we're screening them and I have a full-time research coordinator. So ideally, there are potentially other ways to do this at different time points, but for now, that was the one that we thought was the most feasible.

MEMBER KAPLAN: Right, because the questions of somewhere down the way, decisional regret and things like that get swept into line with how well you've done and you would worry about cognitive dissonance for people who had done very well and they're still upset about the decision they made. Those intervals take -- as I'm assuming as you developed this further, you will help us with the specifications on exactly where you're looking for the magic interval to administer this or say forget it, game over. We can't do this.

DR. SEPUCHA: Yes.

CO-CHAIR STILLE: Okay, Becky.

MEMBER BRADLEY: I agree with everything Sherrie said, but also had some questions about the survey itself and how these questions were derived. They're very different from other types of questions and I'm assuming this survey, it made more sense after hearing the first presentation, but are you anticipating that you would also ask that set of questions? Because there's some relationship to these two studies. So would this be the decision aid survey that would be used to support the previous study? I'm not clear as to how they interrelate.

DR. SEPUCHA: So, you know, really early in Jack's comments he sort of had this idea or proposed this sort of conception of how we do measure the quality of the decision for these areas and so the idea that you want to make sure someone is meaningfully involved in the decision-making process and that was Jack's measure. And that's a little more generic.

We can have the questions kind of generically go across the different topics. But the other two pieces are you need to have somebody who is well informed. So they have to understand the key facts. And we've done a lot of surveys trying to ask patients how informed do they feel? Did they think they understood the benefits and harms in a generic way?

And what we found in that decision study, there was zero correlation between their perceptions of feeling informed and their ability to answer knowledge questions. So we sort of moved forward with saying you know what, in order to actually figure out if someone is informed, we need to ask them knowledge questions. We need to actually figure out what are the key facts that someone reasonably should know who is making this kind of a decision.

For here, it does focus sort of more on the surgical benefits, harms, what's the likely recovery time. Those are the issues that were really important to patients when we were trying to figure out what kind of information do you wish you had known or what information do you think is really important for people making this decision to understand. And so, you know, informed, involved, and then receiving treatments that match their preferences.

And so in this measure, it's just a simple one item which is which treatment did you prefer? Did you prefer to have surgery, non-surgical options, or were you not sure? And we think that the standard for someone who is actually going to have surgery, they should have a clear preference for it. And if you don't, then we shouldn't be operating on them. They should be able to state that. And so that's where this one came up. So these are sort of the decisions-specific items that really have to be tied to a specific topic. And that's why we kind of separated the two different measures, but we often use them together in the same survey.

MEMBER BRADLEY: So patients' input was solicited in the development of the questions?

DR. SEPUCHA: Mm-hmm.

MEMBER BRADLEY: And these were the questions the patient groups said were important to them in terms of knowledge?

DR. SEPUCHA: Mm-hmm.

MEMBER BRADLEY: I guess the exclusions were kind of interesting. So if they failed, three years after surgery, up to three years after surgery, if you couldn't remember a specific statistic about -- I was curious. I probably couldn't remember myself. But it seems like you excluded a lot of people from the study based on the knowledge. But the lack of knowledge -- if they failed the knowledge question, then they weren't included in the study? Is that the way I heard --

DR. SEPUCHA: So do you mean like excluded from the sample, or they just -- so we set a criteria, a threshold, so there's five knowledge questions. And so we set in our threshold you have to have a passing score. So you have to get three or more correct answers to be counted as meeting our criteria for being informed. So I'm not sure, technically, that's not excluding people who didn't --

MEMBER BRADLEY: Okay.

DR. SEPUCHA: -- but it's not counting

them as being well informed.

MEMBER BRADLEY: So, but were they tested at the time that they made the decision for surgery as being well informed or was it only tested at --

DR. SEPUCHA: So we surveyed them up to two years after the surgery in the samples here.

MEMBER BRADLEY: Okay.

DR. SEPUCHA: But I think it gets back to the issue that we were talking about earlier about the timing.

MEMBER BRADLEY: Okay. Thank you.

CO-CHAIR STILLE: Stephen and then Lisa.

MEMBER HOY: Yes, just to highlight, they're not excluded. They're identified as not knowledgeable about their decision. To get excluded, you have to not identify surgery as your preference. So they're just looking at -- right, they're just looking at people who preferred surgery and were they -- and to highlight the difference between this one and the last one, they're looking at their actual knowledge of those decision criterias as identified decision criterias for patients to preferred surgery compared to their process of engagement. So this is a little more direct.

And I have a question, but I'm going to have to come back to it when I remember it.

CO-CHAIR STILLE: Why don't I let Lisa go and then you can remember.

MEMBER SUTER: Again, this is my second disclosure. So I'm involved in the development of a patient reported outcome measure for hip and knee patients undergoing elective procedures. It's not related to this. We're looking at pre- and post-pain and function assessments, but just disclosures.

A couple questions, so first of all, have you -- were surgeons involved in the development of the questions? One of the things I hear working with a technical expert panel and other orthopedic surgeons in the community is the distinction between hip and knee replacement recovery. And that patients -- and I'm also a rheumatologist -- patients recovering from hip surgery actually are far more functional more rapidly. And so I'm surprised to see the time frame for recovery for both of those surgeries is identical. Because when you speak to surgeons and they speak to their patients, they highlight the difference that if you've been through a hip replacement, you're going to expect a different trajectory from the knee replacement.

So one question is how were clinical experts involved in the development of this and the testing of this? 2) reiterating what people have already said, I think if you're going to make this a meaningful measure, you have to bound the time frame that you're collecting this information because two years is, one, it questions what the relevance to clinical care for that particular patient and provider are.

And I guess the last is just thinking about exclusion. I saw you are looking at primary surgeries. What about non-elective surgeries? And are you -- have you considered excluding patients who are -- or at least handling them differently? I know you're not risk adjusting, but a patient who comes in for a fracture, it may be completely unrealistic to expect that person to have any of the criteria for meeting an IPC score in your tool.

DR. SEPUCHA: So the first one, so in terms of the development of the items, we had significant involvement from patients as well as providers. And so the providers spanned from surgeons to primary care docs to physical therapists and nurses. So we had a multi-disciplinary group of providers.

In addition, we also actually have sort of annotated answers to the knowledge questions with relevant citations, so the idea is that these are evidence-based. You did pick up on the one where the evidence is the squishiest, which recovery time in trying to get general ball parks for recovery time that makes sense.

So the idea is not sort of precision with these, with any of the questions, that they know that it's 84 versus 85, it's really generally like do they understand it's a couple of months rather than year or two weeks. Do they understand at least generally? So that's what we've tried to do.

It's definitely a balance. But we have our clinicians sort of review it and the knowledge questions get reviewed every two years just to make sure that nothing is changed such that the knowledge questions would need to change to either be updated or to make sure that they're accurate. So that's the involvement in the knowledge questions.

Bounding the time frame, I do think that's a challenge with the sampling. And I agree. Ideally, we would want to be doing this closer to the -- even before the decision, the surgery was carried out. We've been doing different things. My other hat at MGH, I actually integrate, responsible for integrating shared decision making across the hospital. So we've been working with our orthopedic departments to actually put this as part of their registry. They're already assessing patient-reported outcomes for folks going for hip and knee replacement and adding six questions. It's very reasonable for them to do. They always get a pre-op assessment. So that's the eye toward the future of where we would want to go is to figure out can this get integrated into -- and one of the reasons we picked hip and knee because that's part of the almost eligibility criteria is you have to have a WOMAC score or a patient-reported outcome score of a certain level. And so we thought that this might be a way to integrate that. We just don't have any data on that yet. But that would be the ideal.

For the non-elected surgery, I think we'll probably have to look into it, whether we could actually refine the code so that we could try and remove those. I'm not sure how many primary or hip or knee replacements are non-elective. But that's good. We'll look into that.

CO-CHAIR STILLE: Great. Stephen and Lee and Linda.

MEMBER HOY: Just on that non-elective part, I would still argue that the patient should be informed about what their expectations of recovery and such, so I'm less worried about that.

My question was why the two years? I think you kind of mostly covered that, but it would be nice to see this kind of thing happen before. A good doctor is not somebody who just asks do you understand, it's show me how well you understand. So it would be nice to see it up front and so from what I understand though, it's a sampling issue and a data issue just to get your feet on the ground. But it seems like you answered my question. Thanks.

CO-CHAIR PARTRIDGE: Thank you. I'm going to pick up a little bit on some of the things that Sherrie said, raised. One really goes to the two-year period. And I understand why you did it, but it seems to me that it probably renders this a little less useful.

If you have had the surgery and you're asked three or four or five months later what you knew, it's entirely possible that some of what you know you learned after you had the surgery. It might be particularly true in the case of someone for whom it didn't turn out quite as happily as they had thought. At that point, they start chatting with others and they start reading and they discovered oh, yes, well, I should have known. So they would answer this question correctly three months out, but they wouldn't have answered it correctly just before they actually made the decision.

The second is and I have to tell you I tested this on a sample of one. My husband has had both hips replaced. And he said you know, I would get some of these wrong. And yet, I would get them wrong even after my second surgery because I'm not really sure less than two months or two months to six months which is right.

I don't know that you should flunk my doc because I'm missed less than two months or two to six months. I really knew a lot particularly by the time I had the second.

So his reaction was that perhaps it wasn't quite equitable, again, with number four. If 100, how many will have less hip pain after the surgery, 30, 50? I just have to volunteer.

CO-CHAIR STILLE: Okay, Linda.

MEMBER MELILLO: So I'm going to beat the dead horse. My concern was also with the exclusion and just that even if somebody knew that at three or four months post-surgery, by two years, they've moved on and I would say that even somebody who had a wonderful memory would have a difficult time recalling those kinds of detail so far out. So I'm just wondering if you did any time studies to see at what point people start becoming excluded or also developing the best time frame for which this should be administered.

DR. SEPUCHA: So I think the first

thing, I don't think anybody is excluded. I think the exclusion is maybe the wrong word. So they're not excluded. And yes, I agree, I would rather have the knowledge assessed earlier. So we do have data. I don't have it analyzed for today, but we have data from breast cancer where we had surveyed patients right after their surgery and then a year later. And if we're looking at the site level, so we had four different sites, the average knowledge score at the site level was the same which was surprising. We had actually predicted it would drop. That did mask -- there were some changes on the individual level. Some patients gained knowledge and some patients went from I had an answer to not sure. So we actually didn't find a very big difference up to a year later.

I think we need to do some more work to figure out what the deal is with the surgery for hip and knee. And we have a sample now that was surveyed a week after their visit and again six months after their surgery. So it's not two years and I don't think we'd advocate two years. It was just some sites might need a little extra time to get that number of surveys in.

So I think we can look into getting better data on that and what kind of -- empirically, what impact does that have. I think as we all recognize, you're not going to keep this information forever, but the gist of these answers, is that enough that's going to stick with you? Like, yes, most people actually do better after the surgery. There's very few side effects, but there are some -- but it's not zero. It takes a couple of months to recover. So those kinds of messages, we want to make sure people kind of keep --

CO-CHAIR STILLE: Okay, Sherrie, and then I think we need to start to move to a vote pretty soon.

MEMBER KAPLAN: I guess my question -- I have one comment. And that is the timeliness of these things. Ideally, you'd like to find people just about the time they're making the decision, right? You know, yes, and that's virtually impossible unless we change our data collection methods dramatically.

Then the question becomes timeliness. At what point, if on average your practice doesn't have very well informed patients, does it really matter if you've actually informed them post hoc or before the surgery? It should because you want to have them informed. On the other hand, if you're doing more shared decision making ultimately and you have better-informed patients, then the attribution to the practice of better shared decision making is -- I don't know what that does to you, but that's -- my instant reaction is there's kind of a couple of levels going on here. But before you answer -- address that, you've chosen not to risk adjust and because the Woloshin and this guy named Gigerenzer have issues about numeracy and risk literacy and how crummy that is in the population, even for very well educated people, some of whom are providers, as you found out, have you -- did you consider gist versus numbers? Because Gigerenzer, et al. do that gist thing. I get the general gist of it, but I may get the numbers wrong. Did you weigh the sort of principle I get it that lots of people, not so many people versus actual numbers of people?

DR. SEPUCHA: So we tried to do both, actually. So one of the questions is who is likely to have less pain, people who have surgery, people who have non-surgical options or are they about the same? So that's sort of a gist kind of a question.

But then we also felt actually you need to have a bit of a realistic expectation about what the surgery is going to do, so we also use based on Gigerenzer and that people understand natural frequencies, so we're not using probabilities. We're using natural frequencies. If a 100 people have the surgery, about how many will have less pain after the surgery?

And we've actually tested a lot of

different ways of asking those kinds of questions, whether it's just open-ended, whether you give ranges and this idea of giving -- where we ended up was these sort of -- they're not random numbers, they're 30, 50, 70, 90. Can they understand if it's 90? Actually, almost everybody will have better pain.

So we have tried to balanced that a bit and it's a knowledge question. So we definitely get feedback from some folks who are -- oh, I'm going be tested. Or I don't want my doctor to look bad.

And what we try and do is emphasize even in the questions, this is about how good are we doing a job of getting you the information that you need. And that sort of helps a bit with some of the acceptability and getting patients to fill them out.

And then we also get a lot of notes like I didn't get this -- you know, when we get the surveys back, I didn't know this. Can you give me the answers, you know? Like can somebody send me the answers to these questions?

So that's the other thing we get back when we do the written surveys. So people want this information and they're not always getting it.

CO-CHAIR STILLE: All right, Becky, and then -- I think we'll start voting. We'll have the opportunity for some more specific questions as we go through the metrics.

MEMBER BRADLEY: I just want to clarify. The way I read the exclusions, so if a patient only did not answer three of the five knowledge questions, were they excluded from the survey? Because on the survey I didn't see an answer that said I don't remember. So they either had to answer it wrong or leave it blank. So I guess that just kind of spoke to me in terms of whether or not they were excluded because they weren't knowledgeable or they left it -- just skipped that question.

DR. SEPUCHA: You're right. So we do have -- they have to answer a certain number of the questions in terms of missing data -- in order to get a knowledge score. And so they have to answer at least half of those. So yes, you have to actually answer three of the five.

So the issue about the not sure, so we've done these questions across lots of different topics and lots of different ways and we've often included a response option, I'm not sure. And what we find is that so many people use that that it actually lowers the knowledge. But they know. Like if you asked them to guess, they would guess the right answer, but they kind of use that almost too often and so we were getting sort of artificially low knowledge scores at the end of it.

So what we did is we actually took it out and that did increase the number of missing items but not much. And it ended up, I think, giving us a better sense of what people really did understand.

MEMBER BRADLEY: And I guess my point is that this is new knowledge. It's for patients. And given the age and the population, we know new learning is more difficult to retain than old learning. And so two years after you've been given a survey or a set of information, it may be difficult. They may have gotten it, but they really don't remember it and so it might speak more to the tool or the way the information is provided than to whether or not the provider informed the patient.

DR. SEPUCHA: I think that's definitely a potential challenge with this. You know, that said, the sample that we did have in here where we have four different sites, one of whom had really routine process of getting patients decision aids, we surveyed them about a year after their surgery, and there was a clear signal that the knowledge scores were higher in that site. Whether they would have been even higher if we had surveyed them earlier, I think there's probably a good chance it would have been, but we were still able to actually detect sort of the site that did better or worse a year later.

CO-CHAIR STILLE: Okay, any more burning general questions before we start to progress through. Okay.

Let's start to talk about evidence. Ready to go.

MS. QUINNONEZ: We are now voting on Measure 2958, Informed Patient-Centered Hip and Knee Replacement Surgery. Voting is now open for evidence. Rationale supports the relationship of the health outcome or PRO to at least one health care structure, process, intervention, or service. Option No. 1 is yes, Option No. 2 is no. Looking for one more vote. All votes are in and voting is now closed.

The results are 95 percent voted yes; 5 percent voted no.

CO-CHAIR STILLE: Okay. So we'll talk for a moment about gap. I know there's been a little discussion on this already, but any other comments about performance gap as presented?

Becky, did you have a comment? Okay.

Sam?

MEMBER BIERNER: My question is this was not really tested in a group of patients other than patients at your university hospital, is that right? They don't seem to be very ethnically diverse and kind of match to characteristics of the U.S. population.

DR. SEPUCHA: So there were three different hospitals that we had and one was a community site -- they were sampled. So it was people responding to an ad in the newspaper. So they were potentially treated at different places. So it's not just one center.

MEMBER BIERNER: But your sample is not representative of the U.S. population?

DR. SEPUCHA: Yes, it was not a nationally-representative sample.

MEMBER BIERNER: And you didn't test anything in -- this was only done in English. You didn't have the Spanish version?

DR. SEPUCHA: We didn't have the Spanish version that we've tested. We have a

Spanish version that's available of the measure, but it hasn't been used widely.

CO-CHAIR STILLE: Sherrie.

MEMBER KAPLAN: This is for Sarah. This is another one of these -- Helen. This is for you. Specially designed question for you.

Because you are not phasing yet, the measure is at an early phase of development where there wouldn't be expected to be a ton of data, for example, on geographic diversity or disparities or whatever, help us understand what we're to do with issues of an evidence of performance gap when the measure is in this stage of development?

DR. BURSTIN: Hi, everybody. So it's a great question. And I think this comes up a lot. And I think the question would be if the performance gap can't be provided by the actual measure developer in terms of the work they've done, can you look to the literature is perfectly acceptable as well.

In addition, you can find this matches

the performance gap, it can also just be variation, so you don't just have to have a gap in performance but variation across providers is also applicable, as well as any indication of disparities as another way of looking at fulfilling this particular subcriteria. Does that help?

MEMBER KAPLAN: Follow up on that. Would you consider evidence that it's mutable and responsive to efforts to change the measure like a quality improvement initiative that got implemented as being evidence that there is some room for improvement?

DR. BURSTIN: One would think that logically, if you could move a measure, you've demonstrated there could be variation. It also goes to the conceptual basis for the outcome, which you guys just did. So I would think that would be a logical approach. But also, I would think there's a fair amount in the literature about the variation. We already know about patients' understanding of these conditions and differences and choosing or not choosing to have preference-sensitive conditions. That should also potentially weigh in to the performance gap and variation.

CO-CHAIR STILLE: Great. Any other discussion about gap? Okay. Let's vote.

MS. QUINNONEZ: Voting is now open for importance to measure and report performance gap for Measure 2958, data demonstrated considerable variation of overall less than optimal performance across providers and/or population groups. Option No. 1 is high. Option No. 2 is moderate. Option No. 3 is low. Option No. 4 is insufficient.

Looking for just one more vote. All votes are in. And voting is now closed.

The results are 5 percent voted high; 74 percent voted moderate; 21 percent voted low; and 0 percent for insufficient.

CO-CHAIR STILLE: Okay, now we'll briefly talk about reliability. I know we've already had some discussion, but any other comments about reliability?

MS. SAMPSEL: She struggled with that, did anybody else notice that?

CO-CHAIR STILLE: She tried so hard.

MEMBER KAPLAN: I did. I really struggled. Did I try not to say anything? Yes. So because I raised the issue of if this is going to be used as a practice level variable, if it's going to be used as a practice level measure, right now it doesn't look like the analysis that was done demonstrates that there is inter-class correlations. There is less within them between practice variation.

On the other hand, there's pretty solid evidence that it actually can be used at the patient level. So if the ultimate intent and I don't mean to lead you in a direction here, but if the ultimate intent is to use this as a practice level, do you anticipate that that is the next move for these data or for the evaluations you put this through?

DR. SEPUCHA: So yes. I mean I do

think that's where we're going and getting more data to do that.

I'm wondering if I just clarify like what we had done in that last one because I thought -- so there's one thing where we look at practice level reliability which we might have done the wrong --

MEMBER KAPLAN: Right, but it was a split half reliability and it was within the practice. It was within practice variation, not between practice variation, so in that sense you look for the thumb print across patients within the practice and then compare between practices. And that isn't what was done. But that could be a communication problem with some of the guidance you got. So I'm not sure at this stage in development of this measure that would be as much of a concern.

CO-CHAIR STILLE: I have a similar concern and I don't know since Helen is here maybe we can ask her input about what happens. But in terms of if practices -- if measures are meant to compare practices, but the initial data really doesn't quite have that level of analysis yet kind of where to go.

DR. BURSTIN: I mean, I think basically we need analysis at the level at which the measure is intended. I don't know if there's a higher level of analysis for which the testing done to date would be applicable system or something like that, but that is our requirement, just like, I guess, you talked about probably for the last two days.

MEMBER KAPLAN: Let me follow up on that because of the number of practices, it's not like it's a big available measure where it's been gathered in an environment where you would expect that analysis to be able to be done and if it was not, then okay, with the measure's developer we would ask them and it would be a legitimate question to ask for that. In this case, it's not widely available yet and if it's not being routinely collected, then the sort of between versus within practice variation is not like you have 400 practices to look at, so that's why I'm struggling with we can -- what can we ask legitimately of a measure in this stage of development?

DR. BURSTIN: And this is often the case with new and emerging measures you're not going to have in widespread use, so this is a bit of a Catch-22 because we also find that it is then difficult to get measures into the more widespread use if they're not NQF endorsed. So I think you should look at what you have available to you in terms of the data that have been submitted, knowing it's not going to be for a large number of practices because it is a new measure and see if that's something you could live with as they begin to gather more information as the measure is out.

CO-CHAIR STILLE: Great. Thanks. Thanks for that clarification. Any other comments on reliability? Okay, Desi, let's vote.

MS. QUINNONEZ: Voting is now open for reliability of measure 2958 including precise specifications and testing. Option 1, high; Option 2, moderate; Option 3, low; and Option 4, insufficient.

All votes are in and voting is now closed. Zero percent voted for high; 58 percent voted for moderate; 37 percent voted low; and 5 percent voted insufficient.

CO-CHAIR STILLE: Okay, validity now. Again, we've had considerable discussion about validity, but any other comments that people have thought about? Seeing none, let's vote.

Oh, Sherrie, one more.

MEMBER KAPLAN: Yes, again, well, it wouldn't be -- it's like my bathroom scale analogy. You wouldn't expect -- my bathroom scale is consistently wrong, but you wouldn't expect a measure to be inconsistently correct. So it's really hard given what we just did with the reliability findings to then make the case that then the validity is okay.

On the other hand, having said that, this is one of the few measures I've actually seen that actually does look like it's responsive to efforts to change the measure to quality improvement efforts.

And so in terms of responsiveness and the validity of a measure like this to kind of -- when you actually test it against efforts to improve care, i.e., decision aids that improve shared decision making and the measure response to it, I'm really struggling with how to do this one myself because I feel like this could be one of those very unusual examples of where we would love to have had a different reliability standard, but the validity standard may actually be shifting around on us.

CO-CHAIR STILLE: Thanks. Okay, let's vote.

MS. QUINNONEZ: Voting is now open for validity of measure 2958, including specifications consistent with evidence tests and threats addressed, exclusions, risk adjustment stratification, meaningful differences, comparability multiple specifications, missing data. Option 1 is high. Option 2 is moderate. Option 3 is low. And Option 4 is insufficient.

Waiting for two more votes. All votes are in and voting is now closed. Zero percent voted high; 68 percent voted moderate; 26 percent voted low; and 5 percent voted insufficient.

CO-CHAIR STILLE: Okay, so it passes on validity. Next up, feasibility. We haven't had too much discussion on feasibility, but I don't know how many issues there are.

Becky, you're up first.

MEMBER BRADLEY: Yes, I just wanted to ask about the burden of collecting the data. It seems that this would be done by practices, I guess, but as you all studied this information, how much time was required in collecting the responses and is this something that physicians and clinical practices would have already have people designed to do or is this would be an added burden in terms of feasibility?

DR. SEPUCHA: So I think it depends a lot on the practice. In terms of burden on the patient, the six items were very -- it took a few minutes. It didn't take very long, so it wasn't very burdensome from the patient perspective. But in terms of an additional survey, it would require someone to either mail it out or email it out depending on the method that they're going to be using to do that. So it's -- and maybe a reminder or two. When we've done this, we've had very good response rates with some reminders. So patients are happy to fill this out. I think they like talking about their experiences with these decisions. So we found a pretty good response rate for that. But I think it depends on some practices are going to have more probably resources in place where they're already assessing either patient-reported outcomes, post-surgical, or pre-surgical. So they might have that -- an easier way to sort of integrate these questions into things that they're already doing.

MEMBER BRADLEY: So are you anticipating that there would be a vendor that would collect the information and analyze it at the practice level or are there -- how would the information be used for performance improvement?

DR. SEPUCHA: So I think the practices could collect it themselves and then get that. I don't know if there's a vendor that probably could be doing this as well. I don't think we've figured out exactly yet how that will --

MEMBER BRADLEY: Okay.

CO-CHAIR STILLE: Okay, other feasibility comments? Let's vote.

MS. QUINNONEZ: Voting is now open for feasibility of measure 2958, including data generated during care, electronic sources, and data collection can be implemented. Option 1 is high. Option 2 is moderate. Option 3 is low. And Option 4 is insufficient. Waiting for two more votes.

MS. THEBERGE: David, we need your vote. There it is. Thank you.

MS. QUINNONEZ: All votes are in and voting is now closed. Zero percent voted high; 79 percent voted moderate; 16 voted low; and 5 percent voted insufficient.

CO-CHAIR STILLE: All right, the usability and use. Given that it hasn't been used much yet, but how usable do people feel it is.

Sherrie, did you have a comment?

MEMBER KAPLAN: Yes, I have a question. How do you see this measure being used?

DR. SEPUCHA: So I think similar to what we were thinking about for the measure previously discussed on shared decision making. The idea would be whether it's accountable care organizations. There's lots of things in hip and knee replacements specifically around bundled payments for these things. So thinking about a way to have this incorporated into programs that are looking at the quality of the surgical process of care. And so understanding, you know, not just what the rates of surgery are or the rates of readmissions, but looking at were they the right person in the operating room in the first place, were they well informed, did they have a clear preference for it? So having that kind of bundled in in part of programs that are looking at that.

There are certain folks that have sort of approached us about using them that are payer groups who either have sort of distinction programs or recognition programs for practices or providers who were delivering joint replacement surgeries. And so this could go into that sort of bundled kind of measures that they're going to be using with respect to those programs.

MEMBER KAPLAN: Accountability versus quality improvement?

DR. SEPUCHA: I think it would probably be a little bit of both. You know, right now, it's been only used so far in quality improvement scenarios, but I think there's interest in thinking about it for accountability.

CO-CHAIR STILLE: Jennifer.

MEMBER BRIGHT: So I just want to make a comment that this is where I get really uncomfortable and I know we're in, as a committee, we're in unique, uncharted territories and we're like out there with the machete in the jungle trying to carve a path through, but I get really uncomfortable when we're considering a measure that's really only been looked at within a practice and has high value, I think, from a QI perspective. But then we're immediately jumping and we know the environment into which this gets released is one that's hungry for an answer of what do we pay for. That's when I literally get heart palpitations because I think I want to see measures like this in widespread use.

And I totally get what you're saying, Helen, about there's this doorway that we are helping to facilitate, right? But I get really nervous because we're releasing it into an environment, a payment conversation environment that's not ready for this yet.

And I know we can't talk about guidance, but if we're really only evaluating this based on study data that says it's got high value for quality improvement, we should be evaluating it for quality improvement only and not saying that this has any applicability to looking at accountability. I just really passionately feel like we're in dangerous territory if we say one thing and do another. Sorry.

CO-CHAIR STILLE: Thank you. Sherrie, one more?

MEMBER KAPLAN: It does say on the form under criterion 4 it says "current uses publicly reported, no; current use and accountability or planned use and accountability" and it says no. But then it says yes and no under the current use and accountability. Is that a typo? That's why I asked you the question.

DR. SEPUCHA: Sorry, where?

CO-CHAIR STILLE: The one on the screen just says no.

MEMBER KAPLAN: Well, in the form I'm staring at it's both boxes are checked.

MEMBER BIERNER: You make reference to the Alliance, a regional group in Wisconsin. Are they using this?

DR. SEPUCHA: So they've got a QualityPath program. So they encourage providers or practices to sign up for it and get credit and as part of that, they have to use decision aids and measure decision quality, so use the decision quality instruments from which this measure has been derived. There wasn't a measure available for them to use. They encouraged the providers to use the decision aids.

MEMBER BIERNER: I'm looking at their website document. It's very explicit that they're using it to make decisions about who will -- who is in their network and whether these are preferred providers and things.

CO-CHAIR STILLE: Okay. Lisa?

MEMBER SUTER: So similar to the last measure, have you -- I know patients and providers were involved in the development. Have you gone back to them about how they would use this information and whether they're using this information? I know patients said that they would have liked their doctors to give them this information, but that might actually be problematic after the fact. It could have -- opening Pandora's box downstream saying your doctor really didn't give you a lot of information about the surgery they put you through. So there are some challenges with when you collect this information and how it's integrated into practice. Have you talked to stakeholders, patients, and providers about how they would use it?

DR. SEPUCHA: Yes, we've run a couple of focus groups with patients for hip, knee, and back surgery. And the idea there was to try and figure out what from the surveys what information would they want to understand in order to either evaluate providers or make sense of who they might want to go see.

So the knowledge was pretty straight forward and I think they -- in terms of interpretation of Provider A's average knowledge scores of the patient with such and such and Provider B was something else. I think they really kind of got the sense, oh, they must have -- actually, what they said is they must have a really good nurse who's helping people get information there with that surgeon versus that surgeon. So they really like to see that information.

What was interesting was the feedback that we had given them on the -- how often providers -- you know, patients who had surgery preferred it. And the way they interpreted that is providers who maybe didn't have such a high percentage there so who were actually giving surgery people who didn't have a clear preference for it, they interpreted that as oh, that doctor is really good at talking people into having surgery which is actually in a sense what this is about.

So we have gotten a little bit of feedback of what kind of things, how do they interpret this information, what they would want to do, how would they want to use it. And I think so they liked having that information as well as wanting information, obviously, about sort of the health outcomes, so how do people feel about the surgery if they've had it as another component of it. But we haven't done anything in terms of thinking about how you would release this publicly or what we would do, with the exception of Dartmouth which would publish this on their website. They put their knowledge scores in the distribution of that their patients wanted. But that was just their own breast center website.

CO-CHAIR STILLE: Okay, Linda, real quick. We need to move on.

MEMBER MELILLO: I just have a newbie type question. Is there any way that this committee can put limitations on the use of a measure pending later review?

DR. SEPUCHA: No.

MEMBER MELILLO: Okay.

CO-CHAIR STILLE: Okay, let's vote on usability and use.

MS. QUINNONEZ: Voting is now open for usability and use of measure 2958 including accountability, transparency, improvement, progress demonstrated, benefits outweigh evidence of unintended negative consequences. Option 1 is high. Option 2, moderate. Option 3, low. And Option 4, insufficient information. Just waiting for one more vote.

MS. THEBERGE: David, we need your vote. David, are you on the line? I think he stepped away.

MS. QUINNONEZ: Okay, votes are in and voting is now closed. Zero percent voted high; 50 percent voted moderate; 33 percent voted low; and 17 percent voted insufficient information.

CO-CHAIR STILLE: Okay, so consensus not reached on that. Okay. And then finally, overall, let's vote.

MS. QUINNONEZ: Voting is now open for overall suitability for recommendation for endorsement of measure 2958. Option 1 is yes. Option 2 is no.

MS. THEBERGE: One more check, David, are you there?

MS. QUINNONEZ: Waiting for one more vote. Can everyone resubmit their votes one more time? Thank you. All votes are in. And voting is now closed. Fifty-six percent voted yes and 44 percent voted no.

CO-CHAIR STILLE: So consensus not reached on suitability. Okay. Public comments. Let's open the line.

MS. THEBERGE: Operator, can you open the line for public comments?

OPERATOR: Yes, ma'am. At this time, if you would like to make a comment, please press 1. At this time there are no public comments from the phone lines.

CO-CHAIR STILLE: Okay.

CO-CHAIR PARTRIDGE: We've had a mini-break, right?

CO-CHAIR STILLE: We've had a mini-break. We can --

CO-CHAIR PARTRIDGE: I know we're going to lose people at noon. Do you feel the need for a formal break or can you just all slip out? We want to add one brief item to the agenda. Sarah wants to talk to us a little bit about something else NQF is doing.

So if that's okay, we'll just stay in place and keep going.

CO-CHAIR STILLE: Okay.

CO-CHAIR PARTRIDGE: Sarah, it's yours.

MS. SAMPSEL: And so Helen will be given the opportunity to add to this as well, but based on this conversation and then, you know, kind of it was very interesting to hear all of your comments, we wanted to give you an overview of another project that NQF is working on and really kind of started the conversations between NQF and Jack and Karen and others in this shared decision making space.

So at the end of last year, beginning

of this year, we started a project funded by the Gordon and Betty Moore Foundation looking at the potential application of a national certification program for decision aids. And so the project staff team, as led by Helen and then myself and Andrew Anderson who has conveniently left, but anyways, we are convening a group towards the latter part of this month to talk about what those certification standards might be, as well as what the NQF process could potentially be funding, etcetera. And then part of that conversation as well is starting to talk about how do you translate decision aids and this whole decision making process into performance measures.

And so it was kind of an interesting opportunity when Jack and Karen indicated, and we've been talking to them for a couple of years about as they've been moving their measures through the process. And so wanted, one, you to be aware that a lot of the notes that I've been taking are kind of -- so when we talk to this other panel say these are some of the things and some of the challenges, but also opportunities that we're going to have to start looking at, translating some of these really complex items into performance measures for shared decision making.

So first of all, I just wanted to thank you all because it was very enlightening to hear some of the comments, the questions, and a little bit of the pushback on what we would then turn into a report on developers. If you're thinking about doing these, these are some of your additional considerations.

At the same time, I just wanted to make sure you were all aware of that project going on. You can sign up for project alerts like you do anything else on our web site. But we really find it to be -- I mean at least I do, extremely interesting, extremely novel, but you know, some of these conversations as well, I think, are going to help push us in thinking stronger about it. So it's been great to have the patient input here, but then certainly from you folks who are in practice, going to be trying to turn these around and into use and thinking about how we would then assess accountability and certainly QI in the future.

Helen.

DR. BURSTIN: Just one quick comment in response to Jennifer's comments and others. I think this is very much a sense that we recognize these measures of decision quality are going to be, likely to be very important as we look to the very different payment models that pay for value over volume. And so thinking about how they could be incorporated into future alternative payment models, etcetera, we recognize this is a future tense discussion.

I just want to thank the committee for kind of helping us think these issues through. The panel that will meet later this month will actually put forward a set of standards that we'll put out for comments that we should explicitly share with this committee what those standards for those decision aids would look like because I think a lot of you would find that really interesting. And then actually think of a process of how we'll actually bring those decision aids, evaluate them and try to actually certify ones that meet a standard as the standard is done. And then the logical next step of that if you have decision aids, you logically want some measures that will incorporate the use of those decision aids.

So there is --- this is part of a broader strategy, but thank you for kind of helping us along this journey because this was really informative. And we recognize these issues of it is always difficult when we have new and emerging areas, either when evidence is limited, not that the evidence is contrary, it just isn't there yet. Or when we have new measures where the testing is minimal because, frankly, it's new and it's hard to get people to do a lot of sites of testing of a new measure. So those are issues we will continue to explore.

We've tried multiple ways of bringing in measures in the early stages of development for a quick early check that have generally not worked very well. We had a panel convened to actually help us think through whether we should actually endorse measures for their intended use, should we say this measure is only QI, this measure is only accountability? And ultimately the decision was there's just not enough science on which to base which ones go which direction at this point in time at least. We'll continue to look at it, but appreciate your willingness to look at some very new and I think important concepts that we want to keep moving forward even if not always immediately ready and hopefully the developers can share some additional information back with you as you move forward.

MEMBER KAPLAN: Am I understanding you that you're going to evaluate decision aids and now are you transitioning into evaluating the quality improvement activities versus just the quality assessment?

DR. BURSTIN: No. What came out of our intended use panel that met was that we would not go down a pathway of bringing in measures just for quality improvement. There are thousands of them out there. They don't necessarily need to go through a process like this. And so we decided that wasn't going to happen.

But at the same time there is, some of you may know this history, there's been an international group that has a set of standards called IPDAS around evaluating and setting standards for decision aids. Washington State has been required through legislation to actually begin evaluating decision aids, identifying which ones they're going to, in fact, standardize and use. So we've been building on the Washington State experience and trying to establish a set of national standards we would use to potentially bring in these decision aids so that as part of this value-base purchasing, as part of this sort of brave new world greater patient engagement, there will at least be some sense that these are good decision aids that meet a certain standard. But no so much just for QI, but just broadly we think it's -- just like we bring in cost of research use measures, we don't consider those quality measures for accountability, but they're kind of information building blocks towards a bigger vision of value and so we view similarly this work on decision quality as fitting into that as well, that newer vision.

MEMBER SUTER: Hi, Helen, and I just wanted to ask. I apologize, I stepped out actually to speak with a prior measure developer. Is there a channel for people like this who bring a measure that we're enthusiastic about, but the specifications aren't quite settled and some of the data isn't there yet, although it sounds like she's partnering with a lot of organizations where she might be able to get some of the facility-level data.

Is there an accelerated process for them to come back through or -- I'm just wondering, given that we've put so much effort into thinking through it, is there an accelerated process for something like that?

DR. BURSTIN: So one of the reasons we made the shift to having standing committees was to have the capacity to have a group like you available. So one question, you know, for Karen and her team will be how much they can do in the next couple of months to address those issues? They have until the end of the comment period as you know. If there's anything they could bring back to the committee to move some of those kind of gray zone votes up, that's an option. But we also potentially could convene you -- I'm looking to Elisa who is really the boss of process, to make sure I don't say anything she would -- she'll kick me. But you know, because you already convened as a standing committee, we could potentially do a webinar in X period of time where we would say you've already gone through all these criteria. Let's just go back and look again at the additional testing and maybe just pick up the evaluation from there. Because again, we don't want to have to make these measures sit out there for an eternity waiting for our next call.

So we've been working with CMS to just allow us to have this capacity of you guys are here, you know -- the idea of bringing a measure like that to a new group would make no sense. You've already thought it through. You raised all the issues. Can they address the issues, bring it back to them, and hopefully get it blessed in a shorter time period.

MEMBER BRADLEY: Can I just ask? The scope of the work that you're doing with CMS, is it limited to procedures or does it extend to other informed decisions that patients and families and care givers make like -- because it's kind of counter intuitive to the whole concept of patient choice that CMS has put out there in the past where providers aren't supposed to provide too much information, that patients have the right to just get a list and choose.

So I'm wondering kind of where that overlap is, informed decision versus mission choice?

DR. BURSTIN: Interestingly again for some of these new and innovative areas, we often will go outside of government give us the greatest latitudes to really please the Gordon and Betty Moore Foundation supporting this work.

I will say though that as part of an interesting decision about a year ago, CMS put forward a payment rule, a notice of payment around the use of low dose CT scanning for smokers. And Lisa may know about this because a lot of this is about informed consent of whether you would approach a smoker to have low dose CT scanning. And the way it was put forward by CMS was only if done in the done in the context of use of an approved decision aid, that they specific want for one of these very preference, sensitive decisions that there be that discussion.

I've heard similar discussions around

potentially PSA screening. Again, as you look toward some of these things where the black and white of evidence is fuzzier and it really does come into context of values and preferences, how that all gets incorporated and so we're actually going to work through at the committee, the example of taking the lung CT screening measure and working it through with the committee to think through what that process looks like. But it is definitely a new space and we think it's important. We step into it because we think -- because you look at the broader context of quality and patient-centered quality and particularly it's hard to imagine doing any of that without greater engagement of patients and their preferences.

CO-CHAIR PARTRIDGE: Okay, I'm now going to welcome in person representatives of the University of Colorado Center for Bioethics and Humanities. We apologize for keeping you waiting. Please do join us at the table.

MS. SAMPSEL: And while Matt's coming

up, I'm just going to make a couple of comments and introductions. So Matt and his assistant cohort, Heidi, from University of Colorado as well, had submitted seven measures for maintenance review. And these measures were originally endorsed and approved for endorsement by a Disparities Committee a number of years ago. However, we no longer have the Disparities Standing Committee and the measures best fit in the portfolio for Person and Family Center Care.

However, this was also a recognition of this was a transition process because prior to Matt going to University of Colorado, he was with the AMA, so these measures changed their stewardship. When we received the submission for maintenance, this was a recognition of these might not be ready for kind of this full maintenance review because they hadn't been -- there had been kind of a -- I don't want to call it a dead period, but kind of a sleeping period during a transition. I'll let Matt describe it how he would like to, but let's just say they were sleeping for a bit. Therefore, there wasn't the level of data. There wasn't the level of information that I knew you were all going to look for in the continued and the maintenance of endorsement.

So rather than kind of forcing Matt and Heidi to pull together an incredible amount of data in an extremely short period of time, and through conversations with them, we thought it would be better to bring, have Matt come, given an overview of exactly what they're doing and then look to all of you and the kind of questions that we've had with our other measures over the past few days and few cycles of work to give that feedback on what will you want to see, where are there opportunities so that we can certainly look at these measures in their whole. So that's really kind of the background of why Matt's here and not -- and we're not looking for endorsement of those measures.

DR. WYNIA: Thank you. So it's a pleasure to be here and I expect to learn a lot of really helpful stuff from you today. I should say I spent 18 years at the AMA as a staff person, so when a staff person at an organization like NQF says you know what you ought to do, you ought to come and present in front of the committee and ask them some of these questions. I take that very seriously, so thank you, Sarah, for providing excellent advice in hopefully moving these forward.

So I'm going to give you a little bit of background on this. I was told that the general rules of engagement were that measure developers got two minutes, but that we would get a little bit more than that. And I unfortunately feel like I'm going to need a little more than that because these are unusual measures.

So let me start with where these came from at the American Medical Association. We developed a program called a Ethical Force Program. Actually, John Eisenberg and Linda and Zeke Emanuel and I developed this program long ago. And the idea was to look at ethical performance across the healthcare system. So we felt like organizations like the American Medical Association were pretty good at drawing up ethical guidance for patient-doctor interactions for that dyad. But that then doctors and patients and others were put into environments that made it easier or harder to live up to those ethical standards.

And so we wanted to know are there organizational standards for ethical performance that we could actually measure. And we worked on this with the Hospital Association, the Nurses Association, the Joint Commission, NCQA, CMS, AHRQ, patient organizations, most notably Myrl Weinberg from the -- what's her organization called? National Health Council and so on, so sort of many of the usual suspects.

And the idea was to look across the healthcare system and we looked at specific domains. We looked at privacy and incidentally this was pre-HIPAA, so Mike Fitzmaurice from AHRQ was on the oversight body at the time and took some of the conversations around HIPAA, around privacy back into the conversation around the development of HIPAA for better or worse.

We talked about fair coverage decisions, health system reform and then we started thinking about health disparities and the way in which we frame health disparities and the role that our group could take was to look at effective communication for all patients. We felt like rather than trying to measure cultural competence which was the language everyone seemed to be using at that time, seems a little dated now, but nevertheless, we felt like the way that we should measure whether organizations were doing a good job of creating an environment in which minority patients, people with limited English proficiency, people with low literacy, were getting excellent care would be to look at the communication climate in those organizations.

So in 2006, in collaboration with the American Hospital Association, we did a national project where we traveled to a bunch of hospitals and looked for promising practices in patient-centered communication. This was funded by the Commonwealth Fund. And what came out of it were a set of domains of quality of communication climate in healthcare organizations. We brought those back to the oversight body which again was this 21-member group with stakeholders from across the healthcare system and they voted on these. And it was a fairly elaborate and intense consensus process. We voted on whether each of these domains and then each of the content areas within the domains were universally important, in other words, important to all types of organizations, not to only one type of organization in healthcare, whether they were feasible to achieve, whether excellent performance was realistic to achieve, and whether it would be measurable. And each of these were measured on a ten-point scale, zero to ten. The average had to be seven and if anyone voted three or lower, it was eliminated from consideration.

So the consensus in this instance

included giving everyone on the panel a veto. So this was sort of the first stage of validation in the sense of content validation of what it was we were going to be asking organizations to try and live up to and measure.

In 2010, we had developed a measurement tool kit looking at these nine different domains. And we tested it in a set of hospitals around the country and published those results.

In 2011, we started training consultants to help hospitals, primarily, although some large groups have also used these tools, but mostly it's been hospitals and so we trained consultants to help hospitals use these tools. We did this incidentally, after a period of experimentation where we just gave people the tool, the tools, and said here's how you use them. And it turns out that is an ineffective model of spread. If you just sort of make the tools available which they are, these are publicly available. Anyone can still use them for free, they're online, but the reality is, it is a rare organization that will take a 360-evaluation tool kit off the web and take it home and use it.

So we trained consultants to help people use these tools and the other thing that does is when they work with a consultant, we have some assurance of the quality of data that they are gathering. And so we've created a benchmarking database which only comprises data gathered by organizations working with our authorized consultants. And that allows us to give hospitals these nine domain scores and all of the breakdowns which I'll show you in a moment alongside a benchmark score of what other hospitals using consultants have scored in that domain.

In 2012, the Disparities Panel endorsed seven of our nine domains. And I should say I'll come back to this at the end. We sort of brought it together, we brought it forward as a package to endorse the 360 assessment. And the recommendation at the time was you know, we should look at each of the domain scores separately which would be a little bit like taking the CAHPS survey and breaking it into four or five domains and saying we're going to endorse it four times. And that may or may not have been the right way to do it, but that was the recommendation that we got at the time. So seven of our nine domain scores were endorsed. Two of them were not. And I can talk more about why that was if you'd like.

I will say a point of pride, there were 12 measures endorsed that year for disparities, 7 of which were in our C-CAT tool. So we felt like we had made a good contribution. Unfortunately, we had a number of hospitals that started to use this and then we got a new CEO at the American Medical Association. And with the reorganization of the organization, I was moved into a different area and the continued support for the C-CAT was really brought into question. And as a consequence, people stopped using it.

Not that it wasn't still available to use, but when the consultants realized that the support for this may disappear and/or the benchmarking database may no longer be available, their incentive to get hospitals to use it really disappeared. So we had about a two-year time frame during which almost no one used this. I think we had one assessment in two years. So that's the dormant period that Sarah referred to.

In 2015, I moved to the University of Colorado or began my move to the University of Colorado and started negotiating to transfer ownership of the C-CAT from the AMA to my center at the university. And in January of this year, basically, we relaunched the C-CAT tools.

So these are the nine domains of the C-CAT tool kit and I've put a star next to those that were NQF endorsed. The community engagement domain and the data collection domain are the two that were not endorsed, but as you'll see, performance evaluation, literacy, language services, cross cultural communication, patient engagement, and shared decisions, work force development, and leadership commitment domains were all endorsed in 2012.

This is just the one table that's sort of the most salient perhaps from the validation study. And I'll just highlight one thing because there's a lot of data here. We looked at performance in each domain and whether it correlated with patient belief that they were receiving high quality care; patient trust in the organization in the sense that both the organization would keep things private which is you might think a completely unrelated measure of trust in an organization to the things we're looking at. But it turns out that trust in an organization is pretty strongly correlated with effective communication in almost all of these domains. So a five-point change. This is a 100-point scale. A 5-point change in the health literacy domain corresponds to a 40 percent increase in the odds that patients will believe they're getting high-quality care. Twenty-eight percent increase in odds that they'll say their medical records are kept private by that organization and about a 27 percent reduction in the odds that they would say this organization would hide their mistakes from me. So think of that one, that's obviously reversed -- you reverse code that if you were looking at that positively. But that's again a measure of trust. Will this organization be honest with me if they made a mistake in my care.

So we had quite strong correlations with a couple of exceptions, in particular, that data collection was not related which might not be surprising. Whether the organization is collecting high-quality data on patient race, ethnicity, language, etcetera may not -- there's not a good reason to believe that that would be particularly related.

In any event, these are the tools in the tool kit. So this is a 360 evaluation of the organization. There is a patient survey. There is a staff survey and incidentally, it's an all-staff survey. This is an important point because clinicians have a different perspective on the organization than non-clinical staff do. In fact, non-clinical staff often have perceptions of the organization that are more reflective of where patients are coming from. So if you're talking to housekeeping and engineering and so on, the people who work in the cafeteria, they're going to have a different look at how the organization performs with regard to these issues than clinical staff do.

There is also an executive survey and a workbook that a QI team fills out related to policies and so on. I didn't highlight those because they don't go into the score. They go into the report that you get because they touch on the same issues so they're important in terms of quality improvement, but they don't go into your domain score.

The data collection process is a collaboration between the site, the consultant, and our staff on what the scope is going to be, the time line, preparing surveys for distribution, deciding whose name goes on the cover letter, what logos go on the cover letter, that kind of stuff.

And then all of the analysis of the basic report creation is done by our staff and the consultants don't see this and the hospital doesn't see this. And that's on the cover letter, obviously, to the patients and to the staff.

Where IRBs have looked at this and not always is this required to be reviewed by an IRB, but in some instances it is. Where IRBs have looked at it, the thing they're most worried about is will staff say something in the comment section of their survey that will get back to their boss and end up getting them fired. So it actually has not been the patient survey that's been a concern. The patient survey is entirely anonymous, so it's unlikely that it could be tied back to an individual. But the staff surveys which are also anonymous, but unfortunately, there's a lot of small pieces of organizations where it's hard to maintain anonymity despite efforts.

So the standard report, and I'll show you some pictures from it in the next slide or two, shows the domain scores in each of those nine domains compared to the benchmarks. It also shows executive survey and organizational workbook responses next to the relevant survey items.

I'll just give you a quick example. So you may have a policy on using teach-back as a means of addressing health literacy concerns. You may believe in your QI team that that policy is a pretty good policy and that people are following it. On the staff survey, it may be that many of the staff are unaware that that policy exists and/or that they are not using teach-back in their clinical practice and similarly in the patient survey, you may or may not see that patient's report having ever received a teach-back from their doctor at the institution. So we tie these things together in the report. And we give guidelines for interpretation.

Consultants then often, not always, but often will supplement the standard report with their qualitative findings because they will often have additional services that they'll provide to a hospital, so they may do focus groups. They may do tracer methods. They may do other things which they'll then use to supplement the report and give recommendations for QI or for next steps for the organization.

I want to point out a quirk of this which is that we calculate the overall domain scores by combining the patient responses and the staff responses on similar items. So the teach-back item would be similar for patients and for staff or there would a teach-back item. We had combined those and used that score. And we give equal weight to patient experiences and to staff experiences. So even if you have a lot of patients who will fill out the survey, and relatively fewer staff or vice versa, they weigh equally in the domain score.

We show people in the report both the patient and the staff broken out. But the score that was endorsed last time around is the overall score which is this -- supposed to be this 360 score and I realize that is weird and that you probably don't have any other measures where you're taking data from -- not just multiple items from one survey and combining them, but here we are combining data from two different data sources to create a single score.

CO-CHAIR PARTRIDGE: Matt, I just want to warn you, if you want to leave time for feedback, you're going to lose some of your audience at noon.

DR. WYNIA: Okay, I'll go quickly.

CO-CHAIR PARTRIDGE: I don't mean to ask you to truncate, but I just wanted to warn you.

DR. WYNIA: I got it. I'll skip over the nine domains and how people are using this for QI, but just to say it is useful for QI. This is nine different sites. Just to give you an example of the variation in scores both across sites and within sites, so every site has something that they don't do as well on as they would like. The smallest variation is 20 points which is quite a large variation on these scales.

The report, I'll skip through this. This just shows on the lefthand side how we present the initial data which is to say in this domain, here's your score, here's the component that comes from patients. Here's the component that comes from staff. And on the national benchmark here's the overall, here's the patient, here's the staff components. And then we'll break that out by relevant items from that domain so that you can see specific item scores and see whether there's one or two items that are really dragging down your score.

I mentioned already that we transitioned this to the University of Colorado earlier this year or late last year, early this year. We've done five assessments in the last six months. We have a large health system serving mainly Medicaid populations that is talking about using this for their DSRIP program, so this gets to Jennifer's point about starting to use these types of measures as accountability and payment-related measures which makes me extremely nervous as well. And we've actually got quite a bit of experience using this. But I just worry about making the patient's responses to these items super important because there are so many ways you can game these. This is one of the reasons why we have consultants working on this because it's too easy to game these surveys.

So these are our future directions and I'll just drop to the last one for a moment because it's one of the things I wanted to ask you guys about. We really encourage the sites to over sample minority race, ethnicity, and language populations and so we have a hospital that used this last -- two months ago and I was just on the phone with them last week about their disappointment in their scores. And they did an excellent job of getting non-English language speaking patients to fill out the survey. And that's probably why they didn't score as well on some of these domains. And so they are, in a sense, being punished for doing a better job of collecting data from minority patients. And so we're really struggling with whether we ought to start doing some kind of risk adjustment for the patient populations filling out these surveys.

I'm very familiar with the NQF's background in this area of whether and when and how to use risk adjustment and stratification. We already stratify. So everyone sees the stratified data by race, ethnicity, and so on. But the question is do we also adjust the scores or back weight the scores in some way to adjust for the fact that some sites do a better job than others at recruiting minority patients to complete these surveys.

So I will stop there and look forward to your questions and comments. And if you don't have questions and comments, I have questions for you.

CO-CHAIR PARTRIDGE: Sam.

MEMBER BIERNER: To answer your last -- I would recommend that you not risk adjust for those issues because I think those are important to look at and it's going to vary a lot around the country. Where I came from, we had a 30 -- 50 percent of our patients were -- 30 percent did not speak English. Fifty percent of them were Hispanic and that's why with that other measure I was very concerned because it greatly affects how some of these questionnaires get answered. So I would, I think, stratifying where you can see where it is, but hopefully, there will be a caveat if some larger organization was using it for administrative or payment purposes or where it would cause harm to point that out, that that's something that needs to be looked at directly and an area for improvement obviously.

CO-CHAIR PARTRIDGE: Way down.

MEMBER AVERBECK: Thank you. Beth

Averbeck from HealthPartners in the Twin Cities. Just a couple of comments, one of the -- looks like the outcome based on the CSAT is around trust. I'm wondering -- a couple of comments, wondering have we looked at any kind of quality outcomes and condition outcomes and maybe because it's hospital and most of the outcomes are a process and may not have been correlated as much with outcomes.

As we look to start translating this into an ambulatory world, the types of care and the types of interventions switch to some extent, not necessarily we need to -- obviously, culture of humility adapt how we're approaching this. When I look at hospital care, it's around more rescue care, episodic, acute care. When we look at ambulatory, we're looking at kind of moving toward prevention, chronic care management, and so what I don't know is it the same kind of correlation, the ambulatory world as we would find in the hospital world? It's just a question as you start looking at maybe a broader application of the tool.

DR. WYNIA: So I'm hearing two questions, one of which is what other outcomes can we look at beyond the outcomes that we did in the initial validation study. And the second is what about an application in the ambulatory arena and are they different, are the outcomes different there?

The initial validation studies were in 18 sites, half of which were ambulatory and half of which were hospitals. So we intended this to be used in ambulatory sites to begin with. It has not ended up being used much in ambulatory environments, but I think that may change just because -- and I should say the difference between a hospital assessment and an ambulatory assessment is kind of fuzzy sometimes. We've had sites where an academic medical center uses this across the whole organization. That's both an ambulatory site and a clinic site.

And so I'm not sure how to tease those apart, in particular because a patient receiving this may have been both an in-patient and an outpatient and they are reflecting on their experience over the last month which may have been both. So I find it -- I found it increasingly difficult to say this was an in-patient assessment versus this was an out-patient assessment.

We do have hospitals that ask -- that have added an item on the demographics piece to say at your last visit were you an in-patient or were you ambulatory so that they can tease apart the data in that way.

On the outcomes question I think there are so many things that I would love to have been able to have done with this and we just sort of lost a bit of a window of opportunity because just as it was starting to take off, it was sort of put on hiatus and I feel like in the next year we -- with a little bit of luck we'll have another 50 of these by January if we haven't been contracted.

MEMBER AVERBECK: Just our experience in working in this when we looked at the in-patient, there's less of a gap around the outcome measure because most of those are process measures compared to ambulatory, there's more of a gap in quality in the hospital. There's more of a gap around some of the experience metrics. Anyway, just based on some of the -- on the ground, trying to see what we can do to improve.

DR. WYNIA: Very interesting.

CO-CHAIR PARTRIDGE: Lisa.

MEMBER SUTER: So in the spirit of trying to give you feedback for when you bring this back to us, the things that come up for me thinking through the NQF criteria are in terms of evidence and performance gap, can you demonstrate the use of this is associated with improvements in care. I think that would be given that it's in use.

In terms of scientific acceptability, you raised a huge issue with the fact that you acknowledged that it's gameable. So I'd like to see some evidence for whether or not it gets gamed and how you would address that.

In terms of feasibility and usability, a comment on the burden. It sounds like you acknowledge that it can only be used through certified vendors which I assume is a pretty substantive burden on institutions that are using this. And I guess the last question is sort of is this a measure or is this a service? And if it's a measure, is it one measure, is it nine measures? How are you presenting those to us both from a standpoint of evaluating the evidence. If it's been voted on previously, as seven or nine different assessments, are they usable in isolation or is this really intended as you described as a comprehensive assessment that really shouldn't be broken apart?

I don't quite know what kind of information to ask you to bring back from that, but I am struck by the fact that this appears to be a package as opposed to measure specifications that could then sort of be put out into the public domain or are in the public domain. I don't know the C-CAT well and it seems less similar to some of the other measures that we voted on today.

CO-CHAIR PARTRIDGE: Nicole.

MEMBER FRIEDMAN: I wanted to echo that, too. I was struggling with is this being program or a service versus a measure. And when I was looking at this and some of the -- working with a lot of community-based organizations where -- that serve a larger non-English speaking community, low income, uninsured, under insured. To me, I was looking at this thinking God, this would be a great thing for a large health system's community benefit to purchase or to give on behalf of these social service agencies that are small and sometimes lack some of these tools to self-assess. And maybe I was thinking about it too much in a community benefit form, but I couldn't see -- it was hard for me to differentiate the measure versus service. And that might be a way to avoid looking and penalizing some of the organizations that yes, they're doing a great job in getting the -- not the usual suspects that fill out these surveys, but how do we accelerate that to get more information. Where to do that where there won't be that concern.

DR. WYNIA: I can say with regard to measure versus service, I mean it is both, of course. And the fact that it has the NQF endorsement makes it a much more valuable, both measure and service, to organizations. For what it's worth, people pay attention when NQF says yes, this measure is worthwhile, spending some time and energy collecting.

I do want to make clear, we don't require that you use a vendor. Anyone can do this at any time for free -- anyone who wants can use these measures. It's just that the reality is people don't. They've been available for almost a decade now. And I can't count on one hand of the number of organizations that have picked these up off the web or from publications and contacted me and said we want to do this in our organization and we're very collaborative. I'd love to see more people doing that. We do have -- there's a version of this that's being created for dialysis centers by someone else and I'm totally supportive of that. So we don't require people to use vendors, unless they want their data to go into our benchmarking database in which case we do require them to work with someone who's trained and authorized. But if they want, we will train and authorize someone in their organization. We'll train a new person, but the idea is we need to have some assurance on the quality of the data that are going into our benchmarking database.

Otherwise, -- sorry, the other thing is on the gaming, just to be clear, any survey is gameable, right? That's the problem. We're talking about patient experiences and if we start making patient experiences super important in terms of payment, you'll start getting phone calls from the hospital that sound very much like the phone calls you get from the auto repair guy where it says if you can't give me a ten on this, then don't answer the question, right? Because that's the only thing that matters to us is getting a ten on this. So these are worthless data at that point. And it's just a challenge of using survey data for accountability purposes. They are all gameable.

CO-CHAIR PARTRIDGE: Other suggestions or comments?

Matt, I just have one. I was interested in your comments about the fact that in some situations what you probably are assessing is the entire hospital including its ambulatory care. And that it's difficult to tease it apart. And I was taking that back and thinking well, for now, looking at a set of performance measures that are derived from this information, we're probably going to be asking you to think about teasing that apart or telling us something about why is it okay to lump them all together.

DR. WYNIA: Yes, this is, I would say

the second most frequently asked question by the organizations using this. And that is what's the right unit of analysis? I can say that there isn't one right answer to that because there are organizations that want to say just give me the scores for my emergency department or just give me the scores for the oncology division or give me the scores -- right? And we've had organizations that just did the oncology division. And I think that's perfectly acceptable.

It's also acceptable to do the whole hospital and recognize that the bigger the unit of analysis goes, the less sort of detailed value you get out of that because you know that your organizational climate is made up of a whole bunch of microclimates. But that's also true even within the division of oncology, right? You will have micro climates where one team is much better than another team. And so it's a lumper splitter problem. There are ways to sort of frame this, but I don't think there's one right answer.

One thing we have done, we have -- when working with the consultants, we have told organizations we do not recommend and we will not include your data in the benchmarking database if you lump together five or six hospitals in one assessment. That, we know, is so different from one to the other that we would not -- and I'll tell you why people want to do this. They want to do it because it will be cheaper, right, to work with a consultant to just do one report for all five of our hospitals and our system, as opposed to getting five reports.

CO-CHAIR PARTRIDGE: But if I were running that system I would say I think --

DR. WYNIA: It's almost useless.

CO-CHAIR PARTRIDGE: That's right.

DR. WYNIA: That's exactly right. And that's why we won't facilitate that. I'm not saying it couldn't happen, because someone could still do it, but we won't include those data in the benchmarks and we really strongly discourage people from doing that.

CO-CHAIR PARTRIDGE: Becky.

MEMBER BRADLEY: You mentioned Joint Commission and I'm sure you're familiar with their high reliability organizations. And how does this compare to what Joint Commission is doing? Because most large systems use some accrediting body so as you were describing it, it sounded like sort of like an accreditation process to me as opposed to performance measures that we've typically been --

DR. WYNIA: So Paul Schyve was the chair of our oversight body during the development of this tool and for those who don't know, Paul was at the Joint Commission for many years. And they actually have recommended the use of this as one means of looking at communication in the organization and whether you're meeting their criteria with regard to language services and with regard to addressing health literacy in particular. So in their hospital language and culture report, this was one of the sort of recommended tools for looking at the organization.

The difference is that these are -- you know, we're not using the same methods. We don't use tracer methods, for example. These are surveys of patients and of staff asking about experience of care. So I think these are more analogous to the CAHPS measures than they are to a Joint Commission accreditation.

Incidentally, one of the questions I wanted to ask you all is about an effort that we have in mind to sort of integrate the C-CAT items and the CAHPS items. We went through some rigmarole with CMS early on to make sure that it was okay for organizations to use the C-CAT patient survey because you may know that if you're doing the CAHPS surveys, which everyone is, you can't do other surveys that look similar. And so we had to change some of the items. This was before NQF endorsement. We had to change some of the items in order to make sure that we were not stepping on the toes of the CAHPS survey.

And one of the thoughts we've had is we might be able to do a shorter patient survey than we've got right now if at this time everyone has CAHPS data and we could bring the data in from specific items of the CAHPS survey and use those instead of some of the items we have on our survey and still be able to calculate the domain score. That gets even more complicated now because now we're looking at a response rate for our survey, a response rate for CAHPS, different people responding, right, and trying to -- so it feels like a mess to me statistically and in terms of reliability and so on.

On the other hand, anything to make it easier for people to actually use, you know, makes it easier for people to actually use. Thoughts on that?

MEMBER MELILLO: Yes. I was just wondering if you've tested or intend to test it in other in-patient settings, so post-acute.

DR. WYNIA: Yes, it's been used in a

variety of settings at this point. So we started with this set of testing sites that were half clinics and half hospitals, but subsequently we've had specific units within hospitals do it and we've had entire hospitals do it. It's been mostly that though. It's been either a unit within a hospital, so like the department of medicine or the intensive care units in a hospital.

MEMBER MELILLO: So I was thinking more along the lines of an in-patient rehabilitation facility.

DR. WYNIA: An LTAC, yes.

MEMBER MELILLO: Yes.

DR. WYNIA: I would relish the opportunity to work with LTACs or any amount of SNFs, I'll work with anyone.

MEMBER MELILLO: Okay.

DR. WYNIA: I think it's -- you can pull up the survey. The items are pretty generic, so these are items about effective communication, which is going to be relatively similar across multiple types of organizations.

MEMBER AVERBECK: How many languages do you have available for the patient component of the survey?

DR. WYNIA: That's a really good question. I think we're on the order of 12 or 17. And we'll do additional languages if an organization needs that language. We have a vendor that we use that works --- incidentally, our language translation service will then work with the organization to develop the new language because there are regional differences across the country and so we want to make sure that they do not just sort of a language -- forward translate, back translate kind of thing. They need to actually do a little bit of pilot testing in the local community to make sure the items make sense.

MEMBER AVERBECK: Thank you.

CO-CHAIR PARTRIDGE: Peter.

MEMBER THOMAS: I'm just curious following up on that point. Effective communication is a really important concept in disability law and do you make accommodations for vision impaired, speaking impaired, hearing impaired?

DR. WYNIA: Yes, that's an excellent question. And we have items on the survey about that actually, about the availability of hearing aids and so on within a hospital setting, for example, and glasses within a hospital setting. And what we -- we've tried a couple different things that have not been really successful, so we had for a while, and we still have this, but we just don't use it. We have an iPad version of the survey where the survey is read to you by the iPad in a couple of different languages, I think three different languages.

And we have some -- we've tried to make the survey as easy as possible. It does not have any compound phrasing, unlike the CAHPS. The issues around visually impaired are difficult because the only way we've really got to deal with that right now is a staff person can read the survey to you and help you fill it out. And if so, they are supposed to check a box on the front of the survey to say a staff person helped fill out the survey. So -- because I worry about staff helping people fill out surveys for obvious reasons. But that's how we deal with it right now. We spent $10,000 on that iPad version and no one ever used it. That was eight years ago, seven years ago. It was the first version of the iPad. So it may be that if we were to reinitiate that now, it would be easier because iPads are ubiquitous in hospitals. They were not at that time.

CO-CHAIR STILLE: They're also a lot more reliable, just having done some iPad surveys in our clinical realm with security things and stuff like that. They work. They didn't use to. And I suspect they would be more accepted.

DR. WYNIA: Right, I just want to note that Sherrie Kaplan has not asked me a single question and I'm feeling a little neglected.

(Laughter.)

MEMBER KAPLAN: I have been so good -- all right, I have one question, one quick one because it's time. NIH now is moving towards, they're very nervous about quality improvement activities that are shading over into requiring informed consent. And so because ethics, as you know, is not my field, we just published in American Journal of Bioethics. We asked patients, would you like to be informed about and get consent from different kinds of quality improvement activities? And it turns out the patients are very nervous about sharing their data and other things that hospitals are routinely doing now without their permission.

Have you thought about and in light of our prior shared decision making thing, have you thought about adding a dimension here? I know how standardization works and I wouldn't worry that benchmarking is an issue, but sort of the whole informed consent and shared decision making issue. Have you thought of adding a little --

some stuff in your various different measures that cover that domain? Because this is now beginning to be a little bit dated in terms of its content?

DR. WYNIA: Yes, so we have items on shared decision making about were you involved in the decision? Did people ask for your opinion? Did they respect your opinion, those kinds of items. Did they include you in decision making?

We do not have items about how do you feel about sharing your data, using your data in research or quality improvement work. I was on this Hastings Center group a few years back where we actually came down to say this an ethical, moral responsibility on the part, not only of the organization, but of the patients who come there to allow their data to be used in these ways to improve the quality of care because after all, you're benefitting from the improvements in quality of care. That was not without some controversy, but the idea was, with the right safeguards in place, everyone should allow their information to be used and part of that is driven by the pragmatics of the situation which is it's essentially impossible to segregate out one patient's data when they're doing the QI project in the ICU, and not use just those data because they ask to keep all their stuff private. It's very difficult.

But anyways, it feels to me like that issue would be a separate set of issues and it would be -- they're an interesting set of issues and I know the survey data on this suggests that patients are a lot more reticent -- depending on how you frame the question. They're more reticent to share their information than we would like them to be and probably for reasons that are legitimate.

CO-CHAIR PARTRIDGE: Matt, thank you. I hope this is helpful to you and your team.

DR. WYNIA: Yes.

CO-CHAIR PARTRIDGE: It's certainly been informative for us.

DR. WYNIA: Thanks.

CO-CHAIR PARTRIDGE: So we'll look to seeing you back here.

DR. WYNIA: Yes, we'll keep working with Sarah until we get it right and then we'll bring it back.

CO-CHAIR PARTRIDGE: Break for lunch and then Karen is going to join us, talk a little bit about hospice and it's up to the rest of you all.

MS. SAMPSEL: What we'll do is if everybody will go ahead and grab lunch and take a quick break. Karen Johnson is going to come in. Karen is one of my colleague senior directors and basically, we'll introduce it more, but one of the CAHPS surveys went through a different panel. We're spreading the wealth on some of these fun measures and so Karen just wanted to get some feedback from this group because it has not all gone through the process yet and then for those folks who can stay around, we still need to have kind of a gaps discussion a little bit later.

MEMBER THOMAS: Can I ask a question?

Can we also put on the agenda after lunch if anyone has any comments on any of the measures we looked at yesterday, second thoughts, or just kind of comments that we might want to revisit?

MS. SAMPSEL: Sure. I guess what I should have mentioned there, too, is depending on how many people we have hanging out, I think there are still, and we'd love to get some additional feedback on kind of pushing the field forward on testing as well, which goes into some of those comments and some of the things that Sherrie brought up yesterday. So I think that would work out quite well.

(Whereupon, the above-entitled matter went off the record at 12:06 p.m. and resumed at 12:24 p.m.)

MS. SAMPSEL: So, I think I can actually get started while Kirsten's bringing up the slides. But, basically, we have joining us, Karen Johnson, who is one of my peer senior directors.

And Karen was managing the Palliative

and End of Life Care Standing Committee that met a few weeks ago. And as I mentioned before break, we've started spreading the wealth a little bit about -- on some of the CAHPS measures, some of the experience of care measures, quality of life measures, et cetera.

And putting them into their intended focus areas versus bringing everything to person and family centered care. Part of that is the result of the growing portfolio in person and family centered care.

And just the fact that we're keeping you guys really busy. Which, you know, it's good. It's good job security for me.

But, we think that there are sometimes where some of these measures could really use some input. And perhaps the recommendations from a specialized committee.

So, in starting that transition, one of the first areas was hospice palliative care/end of life care, where Karen's group not only did the hospice CAHPS measures, but also some pain measures. And dealt with some of the same issues that we dealt with yesterday.

But, what -- our intent here then is to create an excuse of information between the committees. And make sure that, you know, the hospice and palliative care -- or the palliative care and end of life committee benefits from some of the input you all may have given.

Because you've looked at numerous CAHPS and experience of care types measures. As well as in the event that if there was another committee looking at something that, you know, just want some of the broader survey or tool-based measure input, we can provide that.

To the same degree, there may be measures that were originally slated to come to this committee, such as there's a renal or dialysis quality of life survey. There are some other surveys that we've started to move into the other portfolios, but we still want to make sure there's an exchange of information.

So, -- and this is for consistency

purposes. But also just really that specialized expertise purposes.

So, that's what we're looking for here.

MS. JOHNSON: So, thank you, Sarah. And if you could just go to the next slide, please.

I will point out that not only is our Palliative Care Committee learning about these kinds of measures, but also I am as well. So this will be the first time I have shepherded one of these kind of measures through.

I'll give a lot of kudos to the CMS folks who filled out their forms. Because they did an excellent job. They made it pretty easy for us to understand what was going on with the various CAHPS measures.

So, I'm just going to walk you through very quickly what the measures were. Give you a little bit of the data that they provided. And then we'll open it up to any discussion and feedback you may want to give.

So, first of all, there were eight PRO-PMs included in this measure. And I know you guys are familiar with the CAHPS methodology. And they used the same scoring methodology, this top-box scoring that the other CAHPS surveys tend to do.

So, they had six measures taken from multiple items, as well as two measures from single items. Those global rating measures. And you can see on this screen, I'm not going to read all of these out to you. But, you can see that there were different numbers of items going into each of these different domains if you will.

Any questions about this? I should mention a couple of extra things. This is, of course, a survey that is fielded to the family member, or really, the primary care giver of hospice decedents. It's fielded about a couple of months after the death of the patient in the hospice. And it is a facility level of analysis.

I think that's the main things. The developer, it was kind of interesting, they talk about several exclusions to the measure. But, when we really did a deep dive on this, they really weren't exclusions to the various measures. They were really things that kept the family from getting the survey in the first place. They were never really in the denominator. And question?

MEMBER BEVANS: Yes, just a quick question. The scoring of the measures, is it -- if it's top-box scoring, but there are multiple items that go into each of the -- so, is it the percent of top-box endorsement?

Let's say on the six items that make up hospice, team, communication, how does that work? Or is it each individual item stands alone?

MS. JOHNSON: Each individual item doesn't stand alone. And I didn't write down the actual scoring methodology.

Let's see, items of never, sometimes, usually, always, is the top-box score. And they -- golly, I wish I could remember exactly what they do.

I think they average it across the four items. And go from there. They did a really nice job of giving us an example of how you calculate it.

And my apologies for not remembering that. I can provide that to you if you want to know.

Any other questions?

(No response.)

MS. JOHNSON: Okay. Let's go to the next slide, please? So, just a few more details about the survey methodology.

As I mentioned, the population is primary care givers of hospice decedents. The measures are all adjusted for mode of administrative, as well as case mix.

And for mode of administration, for these measures, or the survey can be done mail only, telephone only, or some kind of mixed way. And part of the trick there is the individual agency can decide which way they want to do that.

But, mail is cheaper than the other forms. So, there was a little bit of difference. And that's why they needed to adjust for mode.

They do allow sampling for agencies with more than 700 decedent care giver pairs per year. And also, it -- you may or may not know, this survey is already included in the CMS Hospice Quality Reporting Program.

So, hospices are already reporting. And right now, they have, in about, I think they really have only been doing this for about a year.

They only have two quarters worth of data right now. But they have data, so far, on about 70 percent of the survey eligible hospice agencies.

So, it's important to note that hospices with fewer than 50 decedents per year are not required to field the survey. So, that's one way that they work on their small sample size.

And as with the other CAHPS surveys,

they must be fielded by a vendor. So, any question about the survey itself?

(No response.)

MS. JOHNSON: Okay. Let's go to the next slide. So --

CO-CHAIR PARTRIDGE: Just --

MS. JOHNSON: Yes?

CO-CHAIR PARTRIDGE: Karen, just one thing. I'm thinking back, and I was quizzing Sarah too.

And as I remember, we had a brief discussion with CMS about both -- were they asking us to recommend endorsement of the scoring methodology, i.e., top-box, and mode of administration, as well as the measures that were before us. But, were they asking us to endorse how the survey itself was administered, and we explicitly didn't.

MS. JOHNSON: Right.

CO-CHAIR PARTRIDGE: I just --

MS. JOHNSON: Right. And when we're writing this up, we definitely are trying to over and over and over repeat that there are eight PRO-PMs out of this one number, 2651.

So, all this stuff is interesting because it impacts how the measures are calculated. You know, the adjustment, et cetera.

But, we are not endorsing the CAHPS survey or its methodology. You're absolutely correct.

Just to show you here, and I did listen in a little bit to some of your discussion yesterday. And that gave me a little bit of reason to put a couple of these things on the slides.

But, I wanted to show you where we are in terms of performance. And this is just one quarter's worth of data that they had available.

But, you can see here that the scores are pretty high. But, they're not topped out yet, which is encouraging.

I will tell you one of the things that was somewhat encouraging to me with the palliative care project. Several of the other measures that are used in the Hospice Quality Program for CMS, I had actually never seen the results of those.

They're not published yet. They're not, you know, publically reported. And many of those are getting close to being topped out already.

So, it was kind of good to see that, in an odd way, kind of good to see that there's still opportunity for improvement for these measures. So, not only are the means, you know, not quite yet topped out.

But there's also quite a bit of variation across the percentiles there that you can see.

CO-CHAIR STILLE: Just a quick -- oh, sorry, Sherrie first and then I'll.

MEMBER KAPLAN: Karen, I'm kind of shocked that they didn't -- because getting help for symptoms is one of their lowest scores.

And I'm shocked that they didn't do pain management. Because that's one of the

HCAHPS measures. And it's one of the big drivers of the overall rating.

I'm really surprised that they didn't put that in there, you know. It -- again, it's one of the -- it's both -- effective pain management is one of the big drivers of overall rating and recommendations of the care to the hospital to others from HCAHPS.

So, what will -- did they give a rationale for doing that?

MS. JOHNSON: They didn't. Although, and I should have had this up. And I apologize for not having it up. Let me see if I can find it quickly.

What I don't know is the help for symptoms. If the pain question was in there. And I can find that out for you, very quickly if you want to give me just a second.

Sorry, I had no idea what you guys were going to ask. So, I didn't do a very good job of guessing.

So, symptoms. Did your family get as

much help with pain as he or she needed? How often did your family member get the help he or she needed for trouble breathing, for trouble with constipation, and for feelings of anxiety or sadness?

So, that is the three major symptoms on the hospice. So, that's what they did. Great question.

CO-CHAIR STILLE: And now I just had a little point about, you know, when they do all top-box scoring, and your having means, you know, over 80 percent, that means that 80 percent have the highest rating.

And I wonder sometimes, you know, the higher you get, when's sort of the cut off for, you know, clinical, meaningful medicine? And for me it would be if it gets better with intervention.

You know, it has the potential to even get better with intervention. And then I start to wonder, you know, when they get really -- the scores get really high.

So, you know, any evidence about whether there's, you know, improvement projects that are working on that, or something like that. Combined with knowing exactly how they calculated the measure, I think would be helpful.

MS. JOHNSON: I will say that at least a couple of people on the palliative care committee, were a little bit -- they weren't a fan of that top-box scoring. They really thought it was too high a bar.

And you know, so you weigh that with a, you know, is it too hard versus, well, we're looking for great. Right?

CO-CHAIR STILLE: And I guess what I'm thinking is, it's not too hard if you can do something to make it better.

MS. JOHNSON: Yes.

CO-CHAIR STILLE: If you can't do anything to make it better, then it's getting to be too hard.

MS. JOHNSON: Yes. Yes. And you know, it's tricky because it's so new. We don't know if those numbers are going to go up or not.

Again, so many of the other measures that are used in the hospice program right now of the CMS, as I said, some of them are already at the 95 percent. You know, the screening for pain, the screening for dyspnea. All those things.

You know, they're quite -- they're not very exciting measures. You know, I have to say that these are the -- probably some of the more exciting things that we had to talk about in our committee.

And that may be a good or bad thing. I mean, anecdotally they were saying that hospices had actually, you know, even though we're just now seeing data, you know, at this point in time, five years ago when some of them were first starting, the numbers weren't nearly so good.

So, there actually has been improvement. Even though, you know, at our point in time, we're not seeing it.

All right, Jennifer?

MEMBER BRIGHT: Forgive me, because this may be a late question. But I just want to understand who's in this data set?

Is this only for inpatient hospice? Or is this also reflecting delivery of hospice in the home? Or in -- like in the long term care facility where hospice services are brought in?

MS. JOHNSON: Right. It would be any hospice admission. So, hospice is primarily, as you know, provided in the home. I think it's somewhere around 90, 95 percent of hospice care is provided in the home.

So, it doesn't have to be inpatient.

MEMBER BRIGHT: Right. I just wanted to know is that all brought together?

MS. JOHNSON: Yes. It's everybody. Yes. And it's everybody, and there again, I mentioned that whole list of exclusions.

But, obviously if the hospice patient didn't die, they're not given, you know, their family isn't given the survey. And that's a big chunk of hospice enrollees, right?

A lot of people are discharged alive. They don't have decedents less than 18. I'm not quite sure what the thinking of that really is.

I think that -- I don't know if that's a CAHPS usual thing or not. Yes, I'm not sure why.

If the decedent had no caregiver of record, then they can't fill the survey. Or if the caregiver said, don't give me the survey, then they didn't do it.

So, there's those kinds of things that are -- exclude some of those. And that's why I put up that sampling is allowable if you have more then 700 decedent/care giver pairs, if you will.

Any other questions before we go to the next one? Sherrie, you'll enjoy the next slide. Let's look at reliability.

So, this is what they gave us. So, they did give us our data element information. So, how did the instrument itself do using

Cronbach's alpha and then the item total correlations.

And then at the score level for each of the PRO-PMs, they gave us the ICCs. And then did an estimated reliability, assuming that the sample size was 200.

And that of course engendered some discussion. Those reliability estimates are, you know, they're not bad. They're maybe not as high as we might like to see.

Maybe the thing that was a little bit more iffy was the fact that they did it for n equals 200. So, lots of agencies have more then 200 in a year. Right? But, a lot don't. So, there's going to be potentially a small sample effect there.

MEMBER KAPLAN: Okay so I'm confused.

MS. JOHNSON: Okay.

MEMBER KAPLAN: Say it's score -- what does score level mean? Because what we are looking for with these measures that are being used at the facility level, is more between facility variance then within facility variance.

So, at the patient level, it can be reliable. That's, I think that's what they're saying data element level is.

MS. JOHNSON: Yes.

MEMBER KAPLAN: -- is Cronbach's alpha. That's the right internal consistency. Reliability is what you'd look for.

MS. JOHNSON: Right.

MEMBER KAPLAN: Is the patient giving you consistent responses across the questions. But, at the facility level, what you're looking for, is there more between physician variation than within physician variation?

So, if there's a lot of within physician variability, that's trouble at the facility level in terms of reliability. Because then you get a lot of noise.

There is no facility thumbprint across patients in that facility. And that's what it looks like the ICC is telling you.

And Helen, that's kind of a problem.

Should this be -- is this reliable at the facility level?

And if that's what that column is telling you, the third one over, it's not. So --

MS. JOHNSON: So --

MEMBER KAPLAN: And I don't get what estimated reliability for n equals 200 per facility. That's not going to bump the reliability, the ICCs up that much.

MS. JOHNSON: Right. So, I'm just going over my notes here. Let's see. They -- I don't have major notes on what they did for their ICCs and how they did this.

They did use the Spearman-Brown prophecy formula. So, basically what I have here in my notes, they estimated the reliability using Spearman-Brown prophecy formula, assuming 200 surveys were completed in each agency.

MEMBER KAPLAN: But that's what they used for -- that -- it doesn't help you enough if they're using 200 per facility.

And they're using the -- they're not

using some level of testing for is there more between versus within facility variability.

MS. JOHNSON: Yes. So, let me read you what they said.

Inter-unit reliability. Yes, they're not giving us much more than what I've already told you.

And just so you know, Sherrie and I have talked about this a lot. And I'm still trying to learn. I don't -- I completely agree that they should be giving us the between compared to within variation. I believe that's what they think they've done.

MEMBER KAPLAN: I think they did that in the third column. Then the question is, is it reliable at the facility level? See, these measures are great at the patient level. They've all been tested at that level over, and over, and over.

MS. JOHNSON: Right. Right.

MEMBER KAPLAN: We don't need to do a lot of that kind of testing any more with these kinds of measures. Except in a new application like this.

But, when you're -- if you're going to use them to compare facilities, you want the sense that gee, yes, there is a lot more between facility variance than within facility variance. Otherwise, it's great at the patient level. But, it doesn't -- it's not very good as a quality measure of the facility's performance. So, it's not very reliable.

MS. JOHNSON: So, when you're looking at these ICCs, Sherrie, you would interpret this as, that top one for example, 1.3 percent of the variation is between facilities. Is that how you would -- that's how you would actually interpret that number?

MEMBER KAPLAN: See a big number there is .7.

MS. JOHNSON: Right.

MEMBER KAPLAN: That means that there's a lot more -- the score goes up because the numerator is the between facility variance.

The denominator is the between plus within facility variance.

MS. JOHNSON: Right.

MEMBER KAPLAN: So, as the within facility variance goes up, the ICC goes down. So, what you would see, if there's very little between facility variance, and there's a lot of within facility variance, that would give you a low ICC.

And that's what it looks like is going on here.

MS. JOHNSON: so, in terms of them giving the estimate, and apologies, I'm getting a tutorial here, kind of ahead of the game.

DR. BURSTIN: Yes. We can take this up offline.

MS. JOHNSON: Yes. We probably do need to.

CO-CHAIR STILLE: I actually think we all are getting one.

MS. JOHNSON: So, turning the ICC into the estimated reliability with the particular sample size, you don't find that helpful?

Because the committee was looking at that final, you know, when they were making their decisions about recommendations for endorsement or not, they were really looking at that estimated reliability number. That final column, not the third one.

MEMBER KAPLAN: Spearman-Brown prophecy formula tells you how many variables of what -- and with what magnitude of association, plus how many pay -- how many subjects you need to sample that number of measures on with that level of association with each other.

So, if a -- measures are all associated, say, at the level of they're like correlated at .3, say on a, you know, zero to one scale.

So, they're correlated at .3. How many of those measures do you need? And how many subjects per measure do you need to get a stable estimate of whatever you're looking at's performance?

The problem is, I don't know what they did with that final column. I can't -- I don't know what they used in the estimate to get that number.

It would really be bizarre to go from whatever it is that was used to estimate the third column, which I suspect -- they said they had like -- they're at least have a sampling of at least 50, right?

MS. JOHNSON: Right.

MEMBER KAPLAN: And the average number of subjects per whatever was about 200 -- I thought it was about 200.

So, it's really weird to see what's going on. The two columns don't make sense to me.

MS. JOHNSON: Okay. So that's something I'll go back to them and ask them to explain for us. That's very helpful.

Any other questions about this? We're kind of at the max of my understanding of ICCs and reliabilities at this point.

Let's go to the next one. Let me tell you a little bit about what they did for reliability testing.

Basically, and this is probably not unusual for you guys to see. They did some linear regressions with the six multi-item PRO-PMs. So, basically they just regressed those onto the two global PRO-PMs.

And the relationships were statistically significant and decent affect size. And it confirmed our hypothesis of a positive relationship.

They also looked at Pearson correlations between the six multi-item PRO-PMs. The correlations were moderate to large, which is what they expected.

So, from that, they inferred that the different measures reflect what they call unique but related constructs. And between the two of these things they and the committee actually felt that they had demonstrated that the measures are valid.

MEMBER SUTER: Hi Karen. Did they look at either face validity with patients, care givers, providers? Or, you know, any -- or associations with other assessments of hospice quality?

Some of the other measures, whether they, you know, more than three days enrollment before your death? You know, other symptom control outside of the domains that they were assessing?

MS. JOHNSON: Right. They did in terms of demonstrating that the items and the topics were important to caregivers of hospice patients.

They did several focus groups and individual interviews to talk to them. And the results suggested that they find these -- the topic items, communication information, respect, et cetera, to be important facets of high quality hospice care.

So, I think they did demonstrate that the respondents would find these questions meaningful. So, they did that and the committee didn't have any concerns about that.

In terms of did they match it up or do any other analytics to go with the other measures, they didn't talk about doing any of those things. If they did them, they didn't mention them.

So, and I'm trying to remember your idea about the three days is an interesting one. And I don't -- let me look at my exclusion list just to make sure.

They actually don't include decedents who died within 48 hours of admission. So, those really short stays are not included in the measure.

So, that -- and you know, that is kind of limiting the measure. There are a lot of folks who are in hospice for not enough time, right? But, that is one of the exclusions. That might be one of the things that if they did look at shorter stays, they didn't report those. But, the very short stays are not even included.

Sherrie?

MEMBER KAPLAN: Can I ask, Karen, did they go beyond the survey? Because the first, you know, when you correlate these multi-item measures that scales with the two global item measures, you're doing construct validity, but you're within the survey document itself.

MS. JOHNSON: Right.

MEMBER KAPLAN: Did they go beyond the survey document to try and associate its -- these measures with other things we know about the hospice facilities, or no?

MS. JOHNSON: No. No, they didn't. They only did it within the survey.

MEMBER KAPLAN: Because it kind of -- where are they in the development of this thing? Because if -- this is kind of step one.

You want to make sure that the information you're getting from the sub-scales is associated with a validity variable within the survey. Like overall rating, or behavioral intentions.

MS. JOHNSON: Right.

MEMBER KAPLAN: But then you want to go beyond the survey, because you get these methods effect. You know, people are going to give you the same answer within the same data collection method, in this case, the survey.

If you go beyond that method of data collection, then that's where you really get some sense that this is giving you something that's interpretable.

MS. JOHNSON: Right. So, if they did it, they didn't report any of those analyses to us. And I'm about 99 percent sure that they didn't. And I will apologize if they did and I forgot. If it came up during the conversation. I don't think they did.

As to how far along this is, it is fully baked. I mean, hospices are reporting. You know, they are fielding these surveys and reporting on these right now. They didn't mention really what they would be doing in terms of updating things as they went along. They are working, I think they recalibrate some of their risk adjustment models as time, you know, they do that every so often.

But I think that's the only thing that they mentioned so far in how they're doing that. And again, I'm trying to think, I mean, right now there's only, besides the CAHPS measures, there's only seven other measures that are being reported across hospices, at least to CMS. So, across the board. And those are the ones that I pretty much already mentioned. Are you screening for pain? Are you assessing for pain? Are you screening for dyspnea? Are you treating dyspnea?

Let me think. I've got only five minutes, so I better hurry. There's a couple of more that have to do with bowel regimen and spiritual beliefs.

So, they're all pretty basic measures that are pretty much assessed at hospice admission. So, there's not a whole lot there that they can do, again, across the board.

They could do it in a more researchy

kind of way with a smaller sample I would think. But, not with the data that they have in house for all hospices.

Let's go to the final slide. What happened with the committee? You guys definitely asked some questions that our committee didn't.

At the end of the day, our committee was concerned with two of the measures. The treating family member with respect and the getting help for symptoms.

Those were the ones that had the lowest estimated reliability number that N equals 200. So, they pulled those out for separating voting for reliability.

Not for any of the other things. Evidence was the same. Gap was close enough they didn't feel like separating them out there, et cetera.

With the treating the family member with respect, they did not reach consensus on that one. Everything else they did.

So, therefore seven of the eight were

recommended for endorsement. That eighth one, treating family members with respect, we will revisit that after post-comment. And they'll have a chance to vote at that point.

MEMBER MELILLO: So, I maybe showing my statistics weakness as well from grad school, but did they do an item if deleted analysis? Is that the right terminology for the --

MEMBER KAPLAN: What happens to the Cronbach's alpha if the items' deleted from the scale?

MEMBER MELILLO: Yes.

MEMBER KAPLAN: They took a select number of --

MEMBER MELILLO: To select them.

MEMBER KAPLAN: With that number of small, they got some of these are two and three. So, you probably wouldn't --

MEMBER MELILLO: Yes.

MEMBER KAPLAN: Except for the one with six items, you probably wouldn't expect them to do that.

MS. JOHNSON: And again, you know, I got the impression from listening to them that they did a lot of analysis. They didn't show us everything they did.

And that's, you know, not necessarily one of the things that we asked for. So, if they did it, they didn't report that under submission.

A couple of other things that came up, I think, from our committee, is there was a little bit of discomfort actually in the -- Kirsten, can you go back to that very first slide where I show you the -- that second slide, sorry.

The emotional and religious support measure with the three items, some folks thought that that was odd that those tracked together. They really thought emotional should be different then spiritual or religious.

Their Cronbach's alpha apparently supported putting those together. But, that was something that kind of raised flags with our committee.

And I think that's ma -- some of the

major things that our committee talked about. But, do you have other things?

So, it sounds like there's some concerns about the reliability from Sherrie. And Sherrie, is your card up new? Or is that a left over?

MEMBER KAPLAN: It's a left over.

MS. JOHNSON: Okay.

MEMBER KAPLAN: I am trying to keep my mouth shut on this. But, it -- see when a measure is reliable but not valid, it's consistently wrong.

So, just as my bathroom scale I always say is consistently wrong. I love my bathroom scale, it's consistently wrong.

But, in a way that's -- in some ways it's good news because you can recala -- if the thermometer is three degrees off, you know it's three degrees off.

But, you can't have a measure that's inconsistently right. So, when we say something's not reliable, but it is valid, that's weird from a psychometric standpoint. That's really odd.

MS. JOHNSON: Um-hum.

MEMBER KAPLAN: Because -- so, in a way, if you fail the reliability criterion, the validity criterion gets almost uninterpretable.

And so, you know, this is making me get the briggies. Because I'm not sure what to do.

And we did it a couple of times today, where we said something was not reliable, but then it's valid. And I went, how do you do that?

MS. JOHNSON: Right.

MEMBER KAPLAN: So, you know, figuring out how to help advise committees on how to -- when it doesn't pass the reliability issue, what are you going to do then with validity?

And if some of these things have trouble at the facility level, and it's being used at the facility level, -- and then they're not giving it, they're giving us data within the survey, which is still patient level.

And they're not giving us any evidence that -- except for maybe the shared decision making one, where they looked at how mutable it was, responsive it was, efforts to change the scores.

MS. JOHNSON: Um-hum.

MEMBER KAPLAN: You know, that is something that I'm struggling with.

MS. JOHNSON: That's a good point. And you know, I'd be interested maybe offline. Part of me wants to flip our criteria and do validity first. And then talk about reliability.

MEMBER KAPLAN: It won't help any.

MS. JOHNSON: It won't help any? Okay. I thought it might help though.

MEMBER KAPLAN: No, because the issue is still the same. You know, if it's consistently wrong, okay.

MS. JOHNSON: Right.

MEMBER KAPLAN: But it can't be inconsistently right.

MS. JOHNSON: Right. Katherine, did

you have a question?

MEMBER BEVANS: This is actually a question for you. Do you feel like people are getting -- really for the full committee, do you feel like people are getting tripped up in demonstrating that the measure is reliable at the individual level, and therefore moving forward with validity testing kind of at the individual level.

But, really the issue is, they're not demonstrating validity at the aggregate level. And then in the case of like the shared decision making measure, they move forward at the aggregate level with demonstration of validity.

So, is this confusion that people are having, confounded by the need to emphasize reliability testing at the aggregated, meaning at the clinic or provider, or whoever, level?

Because somehow I just don't think that message is still getting to folks. What are -- I mean, do you feel like these two issues are related?

MEMBER KAPLAN: I mean, the issue in reliability terms is, if you're trying to test a measure for use, you know, at the student or the patient level, or whatever, you stay at that level.

If now you're comparing schools or hospitals or whatever, you want evidence that it's -- that you can -- that it's reliable at that level. So, now you've taken it from the patient level, which everybody, you know, knows how to do, to a different level.

And by the way, in the Physician Compare, I just got a note from Physician Compare, they're struggling with the same thing. Because, you know, if you're trying to compare groups of physicians, one to another, then you have to have some confidence that the measure is reliable at the level you're comparing.

So, it changes the -- first of all, it changes the standard. And it changes how you evaluate it.

So, then to go on to say it's valid,

so it may be consistent, but it has trouble with validity, okay we get that. But, it can't be -- you can't have an inconsistently valid item.

MS. JOHNSON: I guess final word. Any other feedback that you would like us to give to our palliative care committee?

MEMBER KAPLAN: Just one?

MS. JOHNSON: Yes.

MEMBER KAPLAN: Can I -- Karen, can I ask that there is a way to do generalized estimation equations that I've you the individual facility score with the standard error.

MS. JOHNSON: Um-hum.

MEMBER KAPLAN: And you plot those. And you see usually what looks like a snake. It will look like a spline.

Can they at least help us? Because what you'll see if it's not reliable, you'll see a big standard error bar. And you'll see you can't tell the difference between any facility in the distribution at all.

MS. JOHNSON: Huh. Okay.

MEMBER KAPLAN: So, if they gave us that, I think, Sarah I gave you one of the papers that we had done with that spline, where it looks like with enough items at the physician level, you can actually distinguish the bottom cortile from the top cortile.

You couldn't distinguish one facility from another. But, you could do that kind of comparison.

MS. JOHNSON: Right.

MEMBER KAPLAN: So, maybe asking them for some generalized estimation equation results, or generate that kind of spline for us. Just to show us the facility and the standard error bars.

MS. JOHNSON: Okay. One of the things that they did give in the time under validity on meaningful difference, and I actually, honestly kind of wondered about this.

But, they did a -- they looked at the hospice national average. And concluded that for all eight of the measures, somewhere in the high 40 percent, there are different from the hospice national average.

So, it's kind of smoke and mirrors a little bit. But, I think that's probably what they're trying to get too at least.

MEMBER KAPLAN: The problem is that again, if you sample a new bunch of patients from that facility, you could get a completely different answer.

So, if there's a lot of within facility variance, then the next time you go back to that facility, you could get a real goof ball answer. And it may not be what they're doing to them, but who they're attracting.

MS. JOHNSON: Um-hum.

MEMBER KAPLAN: That's why this components of variation thing that we've been talking about with Helen is so important.

MS. JOHNSON: Right.

MEMBER KAPLAN: Because if they can give us some sense that the facility level variance is this. Eight percent of the variance is attributable to the facility.

At least you'd have some confidence that, you know, you're not measuring something about the patients. You're measuring something about the facility.

MS. JOHNSON: Right. All right, I will see. Because they have a team. I mean, you know, the CAHPS machine is quite. Yes. Thank you guys.

MS. SAMPSEL: Well, thanks. And I think Kirsten, do you have a couple of next steps slides?

And then what we'll just do is kind of -- what we wanted to do is have a little bit of discussion about any gaps in the measure portfolio.

And you have a list of the full measure portfolio broken down into measures that are currently endorsed, measures that still need to be scheduled for maintenance review, those that lost endorsement.

Just get some general ideas if you think what's still missing in this portfolio.

And that's just -- goes into the measure report.

And is an indication to developers and others that there's still, you know, there's an interest in these measures. And just kind of for future reference for us as well.

And then, you know, as Peter suggested, and I think as we've talked about, you know, work -- I don't want to say we're struggling with, but this committee has a lot of difficult measures all the time.

And so, there are times when some of these measures coming in are getting ahead of, you know, really the ability to translate them into these forms. Translate how we would interpret the criteria, et cetera.

So, I do think some reflections over the past couple of days, and what we can continue to learn and grow, and improve our processes would help as well.

CO-CHAIR PARTRIDGE: And I would like to add on that last point. We specifically, Chris and Sarah and Karen and I, had both email and over the phone, and Sherrie's been part of some of this, a feeling that maybe some of the algorithms we use for the PRO-PMs need to be looked at again.

There -- I forget when that committee report was probably at least -- I should say at least almost three years back.

And we've been struggling with some of these PRO-PMs. And maybe they're still fine.

But, I think there was some sense that maybe -- maybe either we needed to rephrase them so that they were easier for the standing committee members to use. We weren't quite sure.

MS. SAMPSEL: But I think it's broader then that. And I just want to make sure that's out on the table.

It's not just the PRO-PMs. It's any of the tool-based measures. And that's what we're trying to do internally, is make that distinction.

That certainly it happens with the PRO-PMs. That's where, I think, we first encountered our greatest difficulties.

But now that we have the FIM, the CARE, the CoreQ, you know, all of those things. That we really need to make it broader. And so, if some of that feedback is on the table too.

So, just really kind of bring up real quick, next steps. Right now we don't have to have the post-meeting conference call next week.

So, let us internally have our staff meeting tomorrow. Make sure that all business has been covered. But, I think for the most part it is. It has been.

And then, in the meantime, over the next five weeks, staff are writing our draft report. You guys typically get a heads up before it goes out for public comment.

But, then it's out for public comment for four weeks. That's the period that any of those consensus not reached measures that we're working with the developers for those to get additional data back to see if we can bump up those scores at all, because you'll revote on consensus not reached.

The post-draft report comment call, I believe there's a Doodle poll or something out trying to find out when we're going to have that. Because we had some scheduling difficulties.

And then we still have member votes, CSAC review, Board review, and appeals to hopefully finish this all up by the end of the year. That looks like a good Thanksgiving gift.

CO-CHAIR PARTRIDGE: Becky?

MEMBER BRADLEY: Could you explain the endorsement removed? What does that mean when an endorsement is removed?

And is it NQF removes it or the developer removes it? I've just noticed several on this list we're voting on.

MS. SAMPSEL: Yes, I would say that list specifically, a good portion of those measures are AHRQ, they're CAHPS measures. The skil -- both cultural literacy measures out of the CG CAHPS, and then the SNF Chaps.

And so those AHRQ indicated they were

not going to maintain the measures. They're maintaining the tools and the instruments, but no longer the performance measures that go with those.

So basically, in a conversation with them, they indicated they would not maintain the measures. And therefore endorsement was removed.

I think there are two NCQA measures on that list as well. And both of those, I think one they decided just wasn't applicable for measures.

They felt they were more applicable for standards. So, they weren't going to maintain those. And then the other was a health outcome survey measure, the urinary incontinence measures.

NCQA's doing some considerable work on the measure. They've been doing it for a long time. So, it's well past it's maintenance. And they asked for endorsement to be removed while they reconfigure the measure.

CO-CHAIR STILLE: Just a quick

housekeeping thing. When are we going to talk about the related and competing measures?

You're not working, okay. Sounds good.

MS. SAMPSEL: That was a --

CO-CHAIR STILLE: Some other day.

MS. SAMPSEL: Yes. No, that was the thing, you know, we need a little time on that. Because those are the functional status measures.

And I think there are some really big implications that we don't want to hurry the discussions. And especially since with the long term care measures didn't have the same disposition as the SNF measures, which didn't have the same disposition as the IRF measures.

I think we really need to look at that. So, I'm thinking it's going to be post-public comment.

CO-CHAIR STILLE: Okay.

MEMBER HOY: Yes, didn't -- and this may be addressed later then, the UDSMR measures. Some of them were almost related or competing with themselves, right?

In the fact that they were encompassing one another almost. So they were submitting it twice.

So, I guess is that for that conversation then?

MS. SAMPSEL: Yes. So, that's one of my areas of great difficulty. In looking at these, and you know, just in this review for the past two days, the UDSMR measures are related to each other. Same focus area, different population.

However, the SNF group of those measures, at least the self-care and mobility, are competing with some measures that came through last cycle. And then the long term care ones are probably related to some that came through the last cycle.

So, I really just want to take some time to be able to spell that all out and give folks the points of comparison. So that we can really walk through.

The other background for those of you new to the committee is, we had a similar discussion in phase two, and we did not come to consensus on some competing measures. And if there was a best in class.

And so, we really want to think through that. And you know, kind of figure out not only implications, but how that would move forward.

And so, I want to make sure you all have the best information to do that.

MEMBER BEVANS: That's okay, it happens all the time for some reason. I don't know if this is the right time, so just tell me if it isn't.

But, for comments that we may have about the criteria themselves, should we just offer those now?

Something that I noticed in our last review, and especially even more so in this round, is that the criteria somehow don't seem to be encouraging, or the forms or something, don't seem to be encouraging developers to provide a lot of information about the content validity of the measures.

In particular, cognitive testing, qualitative methods that people are using to verify that the measures that are being proposed are understood, meaningful, and I mean the actual items. Meaningful, important and so onto the population in which they're proposing their application.

People seem to talk about that, but sometimes as an afterthought. And I think it's such a critically important component of the overall quality of a measure.

That it's understood. You know that it is meaningful and important to the population of interest.

And so, I guess I'm just wondering, is there a possible -- is there a way to kind of reconsider that validity criteria to call out specifically content validity?

Because, you know, in many cases

actually we saw evidence of people having done some of those procedures. This is a lot of work. It takes -- it's a lot of effort. And if it's missing, it really calls into question the overall validity of the measures.

So, just a thought about that.

MEMBER KAPLAN: In some of these, some of the measures we saw are old in the sense that they've been around a long time. So, patient engagement is a very new construct, concept.

But, it's like -- it's -- some of these measures, not to defend these one way or another, but some of them have been around for -- they were developed a fair amount of time ago.

And then the question is, well then should you not, you know, resurface these measures with a group of folks? You wouldn't expect them to do a whole new content validity with focus groups and reinvent these things.

But, maybe what you're saying is you should check base with some subgroup of folks and make sure they're still relevant to these?

MEMBER BEVANS: I think it depends on who asks with respect to the content validity question. Because there's two thoughts about this one.

Is that the measures are developed and the content validation work is done with one population. And therefore, it's applicable over time and with other populations.

But what we're seeing are measures that are being proposed for use in very specific population. For example, today our chair decision making for seven surgical conditions. And you know, if you asked the FDA, they would say content validation needs to be done at the -- within the specific population that we're recommending a measure to be used.

MEMBER KAPLAN: Where it's more meaning, things like frailty and functional status and those kinds of things that have been around a fair amount of time. Not the newly developed ones at all.

MEMBER BEVANS: No.

CO-CHAIR PARTRIDGE: I'm going to make a pitch for pediatric. Chris and I do it all the time.

CO-CHAIR STILLE: Yes, well me too. Right.

CO-CHAIR PARTRIDGE: But, again at this time --

CO-CHAIR STILLE: As always.

CO-CHAIR PARTRIDGE: We with the 18 and above, and yet we keep thinking that some of them anyway, I guess your comment on the hospice probably would work perfectly well.

CO-CHAIR STILLE: Sure. You know, one of the things that they do I think in like NIH and AHRQ grants and stuff, is you have to provide a rationale if people under 18 are not included.

And I think that would be really helpful. And sometimes the rationale is easy. But other times it's sort of like, yes, well we didn't really want to be bothered or we don't have enough data or something like that.

And that's not necessarily a good

excuse. So, I think even a simple rationale for not including people under 18 would be super helpful.

The other thing is, I was so happy to see the shared decision making measure come up. And there's lots and lots of room for that, for kids as well as adults.

CO-CHAIR PARTRIDGE: I guess I would also, to the extent we can, go back in some of the measures we've endorse. For example, I remember Eric Coleman saying that he was working on a pediatric version of the CTM.

If that's so, we'd love to see it. And we don't know what it would have. Well, okay. But, yes you're right. It has been floated into CAHPS.

There's also a new -- a shorter version of CAHPS floating around, but we've not seen. Do they have to come through with that?

DR. BURSTIN: It's still -- my understanding is, it's still being tested in California. So, again the length of some of these surveys always comes up.

So, there's been a fair amount of testing on a shorter version. We will encourage them to bring it forward.

And to Chris' point about pediatrics, some of our clinical committees, for example, our surgery committee, we always have two -- we've had two pediatric surgeons on there for years.

And they're always the ones who say, so is there any reason for the 18 cut off? And you know, it's interesting because we now have a space on the form where we specifically ask about whether there's any evidence of disparities.

This is an interesting idea to actually ask for justification of the ages at either end. Actually this issue is equally applicable in terms of why is the cutoff at 70 or 80, or you know, maybe just something specific to age and its rationale is something for Karen to kind of build in going forward.

CO-CHAIR STILLE: I think there was one maybe in phase one that we talked about, you know, do you have data for people under 18? And they actually had data for people 14 and up.

Which, you know, was at least a step in the right direction.

CO-CHAIR PARTRIDGE: That was Judy Hibbard.

CO-CHAIR STILLE: Okay.

CO-CHAIR PARTRIDGE: And they didn't have that many people that maybe modified.

CO-CHAIR STILLE: They thought they did. Okay.

CO-CHAIR PARTRIDGE: They were not contributable.

CO-CHAIR STILLE: Okay. All right.

MEMBER THOMAS: Discussion about gaps prompts a question. And this strikes me that this is kind of an investigate or initiated process.

People come to the NQF to get their measures validated. But, does the NQF ever put out what amounts in this research circles to an RFP or an RFA? And request measure sets in certain areas to fill gaps?

DR. BURSTIN: So, it's an interesting question, Peter. There's a couple of different ways we do that.

Certainly at the start of projects, we try to take the gaps you've already identified and put that out there as part of the call for measures. It's often a little bit too late. People are already kind of bring forward things they've worked on for the year.

So, we need to do a better job of that. I don't know if we've shared the document with this group, but we should. We have just a one page around our new strategic plan.

And one of the things that we've identified clearly in there is that NQF will take a very clear role in identifying the highest prioritized measures and the highest prioritized gaps. And we'll develop a set of criteria to use that.

So, we're hoping that will, you know, push people in the directions of filling what we know are national gaps.

And then also again, I don't know if you've talked about it, but we've also formed the measure incubator. Where we have the ability to have folks come forward with ideas for measures and try to hook them up with resources and others.

And interestingly, the measures we've seen most commonly come to the incubator are all PROs. So, we've got a couple of new measures.

We've got some new COPD PRO-based measures being incubated as well as some new work we're just going to launch on PROs for multiple sclerosis and everywhere that we have no quality measures of any kind. So, some real nice opportunities there.

And maybe in some future discussions we can share some of those materials with this group.

MEMBER THOMAS: Well, I just love to -- and on the issue of gaps, I'd love to just put in another plug. I've said this already about the next level of functional measures.

So that they're not so tied to traditional inpatient settings. And that you really get to functional restoration and becoming independent and, you know, all the things that come with that.

And measuring those things are real true outcomes of what people are really looking for after an illness or an injury.

DR. BURSTIN: I should also mention we have a grant with PatientsLikeMe. Some of you may know this group in Cambridge.

So, we have a grant with them to look at what is the pathway towards developing these PRO-based performance measures. With a lot of the emphasis really talking to patients about what they care about and what they think clinicians should be held accountable for.

So, kind of taking it down to more of a really understanding of patients' viewpoints of what's most important here. But also, we'll work with them about thinking about developing.

And one of the ideas is actually just to focus on a set of symptoms for example, instead of functions. Rather then tying it to a clinical area even or a clinical setting at all.

But, great suggestions.

MEMBER MELILLO: So, I'd like to ask the group if -- well, let me back up a step. For the post-acute setting, quality measures and metrics are really exploding at the moment.

And what I saw a little bit of with the EDSMR measures is a tendency to think that you can take it from one setting and apply it to another. And I'm wondering if there's -- if that would be considered a gap?

That we need more measures that are setting specific.

MEMBER BRADLEY: I think part of what we're seeing is that CMS has mandated through the IMPACT Act that we have measures that cut across all the post-acute. And I think that's why we're seeing vendors or developers like UDS come in and -- because that is the directive for post-acute.

So, it would be inconsistent I think to encourage measurement development that's site specific given our mandate. And we know that that's where our funding is going to be in the next few years.

So, it is a dilemma.

MEMBER MELILLO: It is. But just for example, in LTAC environment, we'll have ventilators and measures of that sort. Whereas an IRF will not.

So, I mean, there are definitely major differences in the levels of care. And I understand that the directive for IMPACT is to kind of standardize these cross setting measures.

But, I think that there maybe should be some measures, maybe in addition, that are site specific. So that we can get to really what is the nature of each of those settings. Because they are different.

DR. BURSTIN: I don't think there's anything that precludes that from happing as part of IMPACT. I think there was a desire to have some of that kind of across the settings.

But, there are going to be unique features of an LTAC that are just simply not applicable to a SNF or an IRF. That we want to make sure that patients and others have information that's really relevant to being on an ventilator at a long term acute care hospital.

So, I don't think it's either or. I think it's a good gap. Measures that are really, truly specific to a given setting.

It's sort of different I think then what Peter's saying is, in some ways there is going to be some measures unique to that setting. That we don't want to lose what's special and unique about that setting and capture that as well.

MEMBER FRIEDMAN: I'm really interested in measures around quality for unprecedented partnership between large health systems and community-based agencies.

You know, we're entering a time, I'm speaking for Kaiser, but for large systems where we're seeing a large need to -- for social determinants of health.

And that there's a fear that, you know, we don't want to become a social service agency. But how do we partner with high quality vetted community agencies.

And it's a challenge to know A, what are those quality agencies? And what does good partnership look like?

So, I'd like to see those.

MEMBER KAPLAN: I would make a pitch for expanding the informed and shared decision making group, Helen. To include more broad teaching of how to be an effective consumer of health care.

You know, not just shared decision making which is a fairly narrow thing. But, how to choose and change a provider. How to use the healthcare system to your best advantage. How to be media savvy. How to interpret quality data.

You know, the idea that shared decision making is one area, but it's, you know, there's a lot more to it than that.

MEMBER MELILLO: So, I'd also like to make a pitch for measures that would be across the continuum of care. So, starting in either a physicians's office, primary care, or ER, all the way through to when they finish with services or they're as healed as they can be.

I just think that's something that I haven't seen captured in any particular instruments that are out there.

CO-CHAIR STILLE: And actually Nicole, thanks for bring that up. It sort of rang a bell that we're very good -- well, we're getting good at measuring the medical home.

But really the medical neighborhood which extends beyond the medical home, and even beyond the, you know, medical environment into the community is huge.

And, you know, as we work with patients and families, their care map has, you know, one quarter which is their medical stuff. And then three quarters, which is community, the educational system if you're a kid, you know, family and that kind of thing.

And to the extent, you know, it's going to be baby steps, but to the extent that we can measure connection with important, consistent outside entities that are involved in their care, I think we're going to do better. Maybe the incubator is a good place to start with that.

MEMBER THOMAS: I guess I got a potential gap in process. I'm hearing and getting a lot of chatter on the email about our consideration of the home and community based services measure yesterday.

And how there is an NQF quality committee on home and community-based services. And they're going to be releasing a report next week that provides a quality framework for HCBS quality and highlights the importance of certain domains, et cetera.

So, I guess my question is, to what extent do the committees communicate with each other and one kind of lays out? You know, it just strikes me that maybe there is a little bit of a disconnect there.

Maybe not. I'm just asking the question.

MS. SAMPSEL: No. So, I'll start. So, Karen's one example of coming, for kind of having some of that committee overlap and ensuring that we're exchanging information.

And then HCBS is an example. Disparities is an example. The MAP work that we do is an example for in helping the rule making.

And that's where at the staff level we're ensuring that the other groups and committees, at least their staff, are aware of what we're working on.

And so in this case, with the HCBS measure specifically, that entire team knew about that. And then they were asked to inform their committee of the deliberations as well as the Duals Workgroup, of the deliberations of this committee.

So, it's happening. The one

difference with HCBS is that's not a consensus development project. So, they're not involved in kind of the recommendations of endorsement.

They were a framework project. Which of course, you know, we would eventually have fed into.

So, the communications are happening. We actually encourage them to comment in the pre-evaluation comment. And then they have another opportunity to comment during public comment.

And we'll reach out to the team again to remind them that they can comment to us. And so you don't have to get all of the emails.

Elisa, did you want to add to that?

MS. MUNTHALI: Only to add that the team, the HCBS team is listening in. Some of them were here in the last two days.

And they will be updating the report accordingly based on what happened here.

CO-CHAIR STILLE: Okay. Well, thanks, have a great trip back and thank you, Lee.

MS. SAMPSEL: Well, before you walk

off, thank you Lee. And thank you Chris for leading.

(Applause.)

CO-CHAIR STILLE: We have to open the phone lines for member and public comment. Okay. Let's do that.

MS. THEBERGE: Operator, can you open the line for comment?

OPERATOR: Yes, ma'am. At this time if you would like to make a comment, please press star then the number one.

There are no public comments at this time. I apologize, you do have a public comment from Paulette Niewczyk.

MS. THEBERGE: Okay.

DR. NIEWCZYK: Hello?

MS. THEBERGE: Hi, we can hear you.

DR. NIEWCZYK: Hi, thanks. I just want to remind the committee that the measures, the functional measures that we have put forth for consideration, it was with the intention that they could be used across post-acute care venues.

Since that was not only one of the requirements of the IMPACT Act, but really it was something that was put forth with the PAC-PRD as being, you know, really of major importance.

So, that's something that with that in mind. But also, these measures are just one portion. We certainly understand that there are other measures that would be appropriate for an LTAC, or for a SNF, or for an IRF.

So, this is something that would be just looking at certain functional domains. Whereby other measures could also be collected in those site specific venues.

So, I just want to add that clarification.

MS. THEBERGE: Thank you. Are there any other comments?

OPERATOR: Okay. At this time there are no comments.

MS. THEBERGE: Okay. I think that concludes this meeting.

MS. SAMPSEL: Well, I was just going

to -- Peter, did you have other comments about --

MS. THEBERGE: Thank you very much everyone.

MEMBER CELLA: Bye, bye.

MEMBER THOMAS: Just real quickly. I can't -- I guess I should put it on the table. I kind of feel maybe people in the room will strongly disagree with me. I don't know.

I kind of feel there was a little bit of, maybe we were low on blood sugar yesterday afternoon. But, we kind of treated the UDS measures on long term acute care hospitals pretty negatively.

When I heard some other things today, in particular with respect to the shared decision making data, that was pretty similar to what I heard yesterday afternoon. But, we gave them -- that measure better grades on average.

Now, it's not apples to apples by any means. But, I'm just -- maybe it goes, Sherrie, to your comment about some of the inconsistency with some of what we're -- how we're judging some of these things. So --

MEMBER KAPLAN: Well, I think when a measure is fairly mature, it's reasonable to expect a different level, a different standard from when it's very new. And again, NQF doesn't have yet a phased development kind of standard for us to kind of vary those psychometric, you know, standards to apply.

So, when you see a new measure that's got very limited data because it isn't using traditional data collection methods. It isn't out there.

The lac -- you know, being a little more lenient with the psychometric standards of it seems to me very reasonable when you're first starting off.

On the other hand, when you're using pretty well collected data, or when you're using a standard that -- of measures that have been out there for a fair amount of time and you've got a vast database to work with, then it is a -- in my view, it is reasonable to expect more from those kinds of data.

Especially when there's millions of cases or thousands of cases. And there's, you know, hundreds of facilities and so on.

So, I don't feel badly about holding new -- relatively new measures to a different standard of -- different psychometric standard for myself then I do for measures that have been out there for a fair amount of time and are well used, and we know a lot about how they perform at certain levels.

So, I'm not -- if that was the issue and you know, you sensed something from me, I was reacting to how much we know about the performance of the measures yesterday versus the performance of the measures today, in their phase of development.

MEMBER THOMAS: No, I asked you because I thought that you had picked up some of the same thing with respect to what you had seen -- the issue between reliability and validity.

And what you had -- the point you had

made about how you can't really vote in favor of one, and yet we were. That kind of point.

That's why I --

MEMBER KAPLAN: I didn't, you know, I kind of didn't say that yesterday. I probably should have gone over that.

It is, you know, it is an issue for when you're talking about reliability and validity. It is an important issue to kind of keep in mind.

But again, you know, when you've got limited data to address the reliability issue, you can't really do, you know, what we did either. You know, it doesn't make it any better.

It's just that when the measure isn't -- we don't have much information about how it's performing, it's much harder to make the case that well, here's what we should have seen, and here's what we did see.

And so, and then when we ask the questions, we didn't get the kinds of responses you would have expected for a group that understands what the need is. And then how to respond to it.

Oh, yes, we can do that. We can get those data back to you, no problem. And in fact, in other committees that I've been involved in, when you give that kind of request to the measure developer, some of them just go right back and they'll give you what you want.

MS. SAMPSEL: All right. Any last thoughts? If not, we're really done this time.

I really want to thank everybody. And you know, certainly to the folks who left as well. We appreciate you committing to this. And we know it's a lot of work and a lot of brain function. So, thank you. And I hope everybody has safe travels.

(Whereupon, the above-entitled matter went off the record at 1:36 p.m.)

CERTIFICATE

This is to certify that the foregoing transcript

In the matter of: Person and Family Centered Care

Before: NQF

Date: 06-07-16

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