NATIONAL QUALITY FORUM

+ + + + +

PERSON- AND FAMILY-CENTERED CARE PHASE 2 STANDING COMMITTEE MEETING

+ + + + +

WEDNESDAY
JANUARY 21, 2015

+ + + + +

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, Co-Chair, and Chris Stille, Acting Co-Chair, presiding.

PRESENT:

LEE PARTRIDGE, Co-Chair, National Partnership for Women & Families
CHRIS STILLE, Acting Co-Chair, MD, MPH, FAAP, University of Colorado School of Medicine/Pediatrics University of Colorado School of Medicine & Children's Hospital Colorado
KATHERINE BEVANS, PhD, University of Pennsylvania School of Medicine and Children's Hospital of Philadelphia
SAMUEL BIERNER, MD, UT Southwestern Medical Center
REBECCA BRADLEY, LCSW, National Director of Quality Standards and Case Management HealthSouth Corporation
DAVID CELLA, PhD, Northwestern University
SHARON CROSS, LISW, The Ohio State University Wexner Medical Center
DAWN DOWDING, PhD, RN, Visiting Nurse Service of New York and Columbia University School of Nursing

SHERRIE KAPLAN, PhD, MPH, UC Irvine School of Medicine
CAROL LEVINE, MA, United Hospital Fund
BRIAN LINDBERG, BSW, MMHS, Consumer Coalition for Quality Health Care
SHERRI LOEB, RN, BSN, EMMI Solutions
ANN MONROE, Health Foundation for Western & Central New York
LISA MORRISE, MA, Patient & Family Engagement Affinity Group, National Partnership for Patients
ELIZABETH MORT, MD, MPH, Massachusetts General Hospital/Massachusetts General Physician Organization
ESTEE NEUWIRTH, PhD, Center for Evaluation and Analytics and Care Management Institute Kaiser Permanente
LENARD PARISI, RN, MA, CPHQ, FNAHQ, Metropolitan Jewish Health System
DEBRA SALIBA, MD, MPH, UCLA/JH Borun Center, VA GRECC and Rand Health
PETER THOMAS, JD, Powers, Pyles, Sutter & Verville, PC
CARIN van ZYL, MD, FACEP, Palliative Care, Supportive Medicine City of Hope National Medical Center

NQF STAFF:

HELEN BURSTIN, Chief Scientific Officer
ANN HAMMERSMITH
MARCIA WILSON
NADINE ALLEN
MITRA GHAZINOUR

ALSO PRESENT:

SOPHIA AUTREY *
JEROME CONNOLLY
ANNE DEUTSCH
DANIEL DEUTSCHER
CHRISTINE GOERTZ *
BEN JOHNSTON
MARJORIE KING *
TRACY KLINE *
STEVE LICHTMAN *
JASMINE LARSON *
JANE LUCAS *
TARA McMULLEN
POONAM PARDASANEY

COLLETTE PITZEN *

LINDA RESNIK *

GARY REZEK *

JEANNETTE SHRIFT *

LAURA SMITH

ANITA SOMPLASKY *

MARK WERNEKE

TRACY ZHENG

*** Present via telephone**

C O N T E N T S

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

8:38 a.m.

CO-CHAIRPERSON PARTRIDGE: Good morning everyone. Welcome. I'm delighted to see you here early in the morning. And I'm particularly delighted that so many of us could be here in person. I want to begin our day with a couple of thank-yous. And then I'm going to move on to introduce a couple of new faces around the table. And we'll talk a little bit about the shape of this day and these next two days. Then Ann's going to take us through conflict of interest.

So let me begin first of all by saying thank you to our staff. For those of us, which includes everybody but David, who were here as part of -- participated in phase one. We experimented this time with a different way of presenting the rather large volume of material in a more digestible way for all of us. And I for one think it's a great improvement. So, thank you. I know it represented a tremendous amount

of work on your part because not only did you have to read it all, but you then had to write it up and put it together for all of us. And Sarah's sort of going uh.

And I also want to thank the man to my left, Dr. Chris Stille, who is substituting for Jim Merlino. Jim is back home. His father-in-law is having surgery this week, and not surprisingly, they decided they really needed the family doc in town. So he is not with us. And Chris very graciously has stepped in to co-chair with me these next two days. I want to welcome Charles -- where is, where's -- I'm sorry, not Charles, David, down at the end. David is our new member. And David would you just take a minute and tell us a little bit about yourself?

MEMBER CELLA: Sure, thank you. Good morning everyone. I'm Dave Cella. I'm a Professor of -- in a department called Medical Social Sciences at Northwestern University Medical School.

I, for many years, have developed and

validated outcome questionnaires. Mostly self-report patient-reported outcomes for use in clinical trials and clinical research. And my involvement in performance -- the performance measure arena is a little more recent. Mostly because for about a decade or so I was working with the NIH on some large scale crosscutting measure development work that has you know, possible future application in performance measures. It's sort of under the heading of PROMIS, Patient-Reported Outcome Measurement Information System. So that got me involved into this performance measurement area. But most of my work has been longitudinal clinical research, including clinical trials.

CO-CHAIRPERSON PARTRIDGE: Wonderful. Thank you. And it's going to be a valuable addition.

DR. BURSTIN: And if I can supplement, David also wrote our Commission paper for all our -- for our PRO work. So quite steeped in this phase. So thank you for that contribution. It

was really quite significant to this work.

CO-CHAIRPERSON PARTRIDGE: And Ann Monroe is joining us in person for the first time. Ann and I served on CSAC together. And I'm delighted that she's on this Committee. And I'm especially delighted that she's able to be with us in person. So Ann, why don't you also tell us a little bit about you?

MEMBER MONROE: Thank you. I'm Ann Monroe and I'm the President of the Health Foundation for Western and Central New York. And I wasn't here for the first session because I was a patient. I've had three surgeries on my leg. It's been eight weeks in rehab. And trust me, this is very relevant, although a little overwhelming to what I'm doing. So I'm very glad to be back and to be participating fully.

CO-CHAIRPERSON PARTRIDGE: We're certainly glad to have you here. And another face who's unfamiliar is Suzanne Theberge. Who is a veteran NQF staffer but is now with this Committee. Which just delights me.

Suzanne and I worked together on an earlier standing committee. And as I told you when I learned she was joining us, I think you will all enjoy working with her a lot. Suzanne also tells me she's a new mom. And if you get a chance to see that picture of that little girl, you'll just know it must be very hard for her to be here with your daughter back home.

All right. We are going to -- Chris and I are going to alternate sitting in the chair today. He's going to lead our first session this morning so that I can participate a little more fully in the discussion of that block of measures which were ones that I personally was assigned to review.

We slotted it somewhere between 15 and 20 minutes a measure. But I think, probably, we will not adhere rigidly to that time frame. Particularly, since as we know, some of these measures overlap there. The first block is very similar in structure at different body parts, but a lot of the same issues come up.

So I think we will try to stay within the broad parameters of our time frame. But not necessarily strictly to 15 to 20 minutes. Chris and I are going to keep tabs on the time and we'll try very hard to keep us on track. We also want to leave ample time for some of the discussion items that are at the end of the agenda tomorrow. A few of us, I know Chris for one has to leave in time to catch a plane. So we may move that -- one of those discussions up into today if the scheduling works out to allow that.

We will, I promise you end on time both days. And I always believe in trying not to work through lunch. It's a time for us to get a chance to get acquainted a little better. To check our emails, et cetera. And I -- we may not be able to stick to that, but I'm going to try. So with that, I am going to turn it over to Ann who's going to take us through conflict of interest.

MS. HAMMERSMITH: Thank you Lee. Good morning everyone, I'm Ann Hammersmith. I'm NQF's

General Counsel. And as Lee said, I'm going to take you through the conflict of interest disclosure. We'll combine that with introductions because we find that that saves a little time. I'm going to go through a few points regarding the disclosures and then we'll go ahead and go around the table.

If you recall, you received a rather lengthy form from us where we asked you for information about your professional activities, such as research grants that you may have received, consulting that you do and so on. What we're looking for you to do here today is to go around the table and disclose things that you believe are relevant to your participation on this Committee.

So in other words, please don't summarize your resume. I know that you have a very full agenda, so we don't want any resume summarizing. But disclose things that are relevant to the subject matter of the Committee and the work that the Committee is doing. For

example, we are particularly interested in your disclosure of research activity, grants, and consulting is relevant to the work of the Committee. Anything else that you think is appropriate to disclose, such as you may have served on a committee that has something to do with the subject matter.

And one of the things that's a little bit different about NQF's disclosure of interest and conflict of interest process is we don't only look at financial conflicts of interest. So, in other words, you may have done something as a volunteer that may be relevant to your work on the Committee and that may be something that we would expect you to disclose.

A few reminders, you sit on the Committee as an individual. Sometimes people get tripped up by that a little bit and they'll say I'm Susie Jones, and I'm here representing the American Academy of fill in the blank with whatever the subject matter is. Actually you don't represent any organization. You don't

represent your employer. You don't represent any organization you may be associated with. You don't represent any group that may have nominated you for service on this Committee.

You're sitting as an individual. You're on the Committee because you are an expert. And we're looking for your expert assessments and opinions. So with that, let's go around the table, tell us who you are, who you're with. If you have anything you would like to disclose. Another reminder, just because you disclosed does not mean you have a conflict. Part of the idea of this is to be open and transparent, know where everyone is coming from. So, let's start with the Co-Chairs. I always start with the Co-Chairs.

CO-CHAIRMAN STILLE: There we are. Hi, I'm Chris Stille. I'm a general pediatrician and head of the Division of General Pediatrics at the University of Colorado School of Medicine and Children's Hospital, Colorado.

My work has a lot to do with the

patient-centered medical home and development of measures for coordination of care within the patient-centered medical home. I'm also at the American Academy of Pediatrics. I sit on the Committee of Children with Disabilities. Having said that, I don't believe I have anything to disclose related to these measures. Really nothing at all. I'm happy to discuss them.

CO-CHAIRPERSON PARTRIDGE: And I'm Lee Partridge. I'm Senior Health Policy Advisor at the National Partnership for Women and Families. I am also a current member of CSAC here at the National Quality Forum. I have -- I am a colleague of Chris' on the American Academy of Pediatrics' Patient-Centered Medical Home Committee. So in that role we work actively to encourage pediatricians to become qualified medical homes. But I don't think it has any conflict issue.

I also serve on the Clinical Programs Committee of the National Committee for Quality Assurance, which does not develop measures.

That's the standing committee of NCQA that approves the recognition tools for among other things, being recognized as a medical home. And I have nothing to disclose.

MEMBER VAN ZYL: Hi everyone, I'm Carin Van Zyl. I'm a Palliative Medicine Physician at City of Home National Medical Center. Sadly I do not have any grants, research or consulting fees that would put me in conflict with any of these measures.

MEMBER BEVANS: Good morning everyone. I'm Katherine Bevans from the Children's Hospital Philadelphia. I'm an Assistant Research Professor there. I'm a health outcomes researcher. I have received funding from both the National Institutes of Health as well as Patient-Reported Outcome -- I'm sorry, the PCORI, Patient-Centered Outcomes Research Institute, to develop PROMIS measures, the Patient-Reported Outcome Measurement Information System. Measures are sort of generic patient-reported outcome measures. However, content-wise, I don't think

that poses a conflict. Thank you.

MEMBER DOWDING: Hi, I'm Dawn Dowding from the Visiting Nurse Services of New York and Columbia University School of Nursing. And I am not involved in any research that relates to the content measures we're discussing today.

MEMBER MONROE: Ann Monroe again. I'm not involved in any research on measures. Although I do sit on CSAC here at National Quality Forum. I also am on the Governor's Medicaid Redesign Team in New York, which means that we'll be using measures. So I don't know if that counts. But --

MS. HAMMERSMITH: It's not a conflict. But thanks for disclosing it.

MEMBER MONROE: Oh, okay. All right.

MEMBER NEUWIRTH: Hi, I'm Estee Neuwirth with Kaiser Permanente. And I'm the Director of Evaluation and the Care Management Institute at Kaiser Permanente in its national offices. And I mostly do very applied studies and research to understand opportunities to

improve care and spread leading practices. I don't have any research or consulting related to these measures. Thank you.

MEMBER THOMAS: Hi, I'm Peter Thomas. I'm with Powers, Pyles, Sutter & Verville, it's a law firm here in town. And I do healthcare law and represent a lot of clients on rehabilitation and disability issues. I have never participated in the development, I've never assisted in the development of a measurement tool. But I do represent a number of clients, and my firm represents a number of clients that are engaged in this work.

I don't advocate on behalf of particular tools. There is one that I disclosed involving a measure that's not on the table today that I've had some involvement in. But ultimately all of this is very familiar to me and some of the organizations are very familiar to me, and in fact I represent some of them. I'm happy to name names if you'd like me to. But I've never done any development or advocacy on

behalf of any given measure.

MEMBER LINDBERG: Good morning. My name is Brian Lindberg and I'm the Executive Director of the Consumer Coalition for Quality Healthcare. I also work with a number of nonprofits generally who don't have Washington offices, on policy development. And I have no conflicts related to the measures. Thank you.

MEMBER CELLA: Hi again. Dave Cella, Professor of Medical Social Sciences at Northwestern. Helen mentioned the white paper that I led the writing of on patient-reported outcome performance measures that I submitted to NQF now a year and a half ago or so.

And I am involve -- and I mentioned PROMIS in my earlier introduction, which I've been involved in for 10 or 11 years. More closely related but not, I don't think related enough to be a conflict, but I'll let you know, I'm involved with two current projects, both funded by PCORI to develop performance measures. One of them is to develop extensions of PROMIS

generic item banks into knee replacement candidates and heart failure candidates. That's led by Dartmouth and I'm a co-investigator.

The other one is led by Allen Heinemann at the Rehab Institute of Chicago to develop performance measures for acute care rehabilitation facilities. But I don't see anything on this list today that poses a conflict that I sense.

MEMBER SALIBA: I'm Debra Saliba. I am a Professor of Medicine at UCLA and the Veterans Administration in Los Angeles. I also work at the Rand Corporation. I'm a geriatrician. I do health services research and I've been funded by multiple organizations including NIH, CMS, ASPE, AHRQ. And I currently am going to recuse myself from one of the measures, 0688, because I was a member of an expert panel that gave feedback to the measure developers on that particular measure.

I'm on the Board of Directors for the American Geriatrics Society, a nonprofit

organization that advocates for patients and providers of older adults. And I'm also on the NQF's post-acute care and long term care expert panel. And I sit on the five-star TEP for CMS to advise them on their quality of metrics that are part of their Compare.

MEMBER MORT: Good morning. My name is Liz Mort and I'm an Internist at Mass General in Boston at Partners Healthcare. And I'm the Senior Vice President for Quality and Safety.

I have no conflicts, but I have had long-standing interest in patient-reported outcomes, having done research in my fellowship in 1990. But I'm not sure that's relevant in 2015. It does take a long time for these things to translate, I was telling Estee in the elevator. I'm very interested in the use and I promote the use of PROMs in our hospital in the system. But I do not have any conflicts. I also participated in the NQF program on PROS about a year and a half ago.

MEMBER PARISI: Good morning. My name

is Len Parisi. I'm the Vice President of Quality Management for Metropolitan Jewish Health System in New York. We are a post-acute care provider and two managed care plans, Medicaid and Medicare.

I don't perceive any conflicts of interest. I am fortunate enough to have -- to use many of these measures in our post-acute care work both from a long term care and home health. I also had the opportunity to be a beta test site for the OASIS outcomes in the late '90s. So I'm looking forward to the discussion today. I am the immediate past President of the National Association for Healthcare Quality. And recently appointed to the Joint Commission on Standard and Survey Procedures Committee.

MEMBER BRADLEY: Good morning. I'm Becky Bradley. I'm a social worker, so I'm not a statistician. So I look at these measures that we'll be discussing today more from how they can be used for quality. I am the National Director for Quality for HealthSouth Corporation, which is

an inpatient rehabilitation company. We also have some home health agencies. So I'm very interested in these measures.

I've been in this field since 1980 and watched many of these measures be developed from the sidelines, not because I've been involved in the research. But we do use these measures and I'm very familiar with many of them. I did disclose that I sit on a product advisory committee for UDS which is one of the measurement developers that will be presenting today.

MEMBER KAPLAN: I'm Sherrie Kaplan. I'm a psychometrician by training. Which I always joke that my mother has no idea what I do for a living, and she still does not. So I am a Professor of Medicine. I'm Assistant Vice Chancellor for Healthcare Measurement and Evaluation at UC Irvine. I currently co-chair the Admissions-Readmissions Committee at NQF. I'm also just about to sign a consultant agreement with NQF for helping out I guess in general about measurement.

I have a grant from PCORI to enhance and develop a measure -- a self-reported measure for children ages four to 12, animated touchscreen-based measure of children's general function status, enhanced to do a module on perioperative anxiety and pain management. That's ongoing.

And I also have been involved since my Rand and UCLA days with the Total Illness Burden Index, a patient-reported review of systems that can be scored to summarize severity and complexity of illness. And we're currently have a contract with Eli Lilly to develop that for a priori stratification of randomized trials. I don't think I have any conflicts around either of those issues in these measures.

MEMBER CROSS: My name is Sharon Cross. I am part of the Patient Experience Department at the Ohio State University Wexner Medical Center. My background is in oncology social work. So like Rebecca, I am not a specialist in statistician work. So look forward

to hearing the discussions from everyone else. But I do feel like I do a great job of representing the patient and family needs. I've been a chronic patient myself for many, many years. So I come from that background.

I do have a consulting job on the side with PCORI, but it is specifically in helping train reviewers who are looking at grants. So I have no impact on where the money goes or who gets selected for a grant or not. So I don't believe I have any disclosures that -- or any interest that I -- or conflict of interest that I need to disclose. Thank you.

MEMBER LOEB: hi, I'm Sherri Loeb. I am a nurse, have been for a long time. As of next week I will be working for Advocate Healthcare in the Chicago suburb taking care of Alzheimer's and dementia patients. So nothing really that fits with these measures per se of the -- but what brought me here was personal experience and advocacy that we need representing all patients. So I don't feel I have any

conflicts. But you know, really feel that this is critically important.

MEMBER MORRISE: My name is Lisa Morrise and I am a broadcaster. I have a background in media management and teach media management for Brigham Young University. I'm also a mom. Probably my most important thing. And my daughter was born unable to breathe or swallow and we've had -- just had surgery number 44. She's almost 22.

So I got involved in patient advocacy and became a specialist in patient and family centered care in patient and family advisory councils. I'm doing a webinar next month for the Institute for Patient- and Family-Centered Care. It will be my fourth time doing this particular webinar in how to train patient advisors.

And I work with Consumers Advancing Patient Safety and Marty Hatlie and Natasha Washington. And I'm developing webinars for them around patient- and family-centered care and patient advocacy. I don't think that there is a

conflict there. I don't work with measures anywhere or research. I'm an advocate.

MEMBER BIERNER: I'm Sam Bierner from Dallas, Texas. I'm a physiatrist, a specialist in physical medicine and rehabilitation. I work in an academic institution where I'm involved in developing clinical guidelines for treatment of back pain. And also involved in inpatient rehabilitation and quality assurance. I don't have any grants currently. And I have no conflicts of interest otherwise.

MS. HAMMERSMITH: Okay. Thank you for those disclosures. Based on the disclosures this morning, do any of you have anything that you want to discuss with each other? Any questions of each other? Okay. Before I leave you, just one more reminder. In order for a conflict of interest process to really work well, we rely on each of you to participate actively. What that means is, if you think you have a conflict of interest during the meeting, please speak up right away.

If you think somebody else has a conflict of interest, or if they are acting in a biased manner, we'd also like you to speak up in real time. You're always welcome to do that during the meeting itself. If you'd rather not handle it that way, you can approach your co-chairs who will go to NQF staff, or you can approach NQF staff directly. What we don't what you to do is to sit there in silence and then six months later say you know, not quite sure if that was okay. So, that's my final reminder. And have a good meeting.

CO-CHAIRPERSON PARTRIDGE: Thank you Ann. And I'm now going to turn to Helen.

DR. BURSTIN: Good morning everybody. Again, thank you for those introductions. I am so struck by the breadth and depth of this Committee. It's really quite staggering. I can't imagine a better group to evaluate the measures before us today. And your work won't be done, because there's many more in the queue to follow in this particular space as it grows and

grows.

So I'm delighted. I mainly just wanted to take a chance to say good morning, but actually mainly to introduce Marsha Wilson, who has just joined NQF a week ago as the new Senior Vice President for Quality Measurement here. As some of you know, that was the job I had had, and I am now the Chief Scientific Officer. So we'll be working really closely together.

But Marsha joins us from years of helping to lead Aligning Forces for Quality, a community-based initiative with the Robert Wood Johnson Foundation. So brings a wealth of community based and implementation experience that I know we sorely lack. So we're just delighted to have her. You'll see lots of her at these meetings. I'll still be here to help with any of the science issues. But you know, really she will be leading this department. So I just wanted to add a welcome to her.

CO-CHAIRPERSON PARTRIDGE: Welcome to Marcia also. And welcome to the rest of the NQF

staff, Mitra, Nadine, for keeping us in line and keeping us informed. And always being so nice about it. Sarah?

MS. SAMPSEL: Yes. So, good morning everybody. And I do want to take just a real quick opportunity to have Mitra, Nadine and Suzanne introduce themselves as well. I started with this group at your last meeting in phase one, and kind of shadowing Karen Pace and Lorelei so I could pick up the work since they've subsequently left the organization.

I'm a consultant to NQF with years of measure development and implementation history. But you know, just want to reflect that it really is the staff team that has prepared these documents, as well as prepared and the logistics for today as well. And you know, as Lee mentioned earlier, we have changed the process a little bit. And we warned you of that at the end of the last phase, regarding removal of the workgroup meetings and starting to do a staff review.

And so it will be really interesting to us if you have any feedback on what, you know, what different things we could do for that staff review. We were really challenged by this measurement set due to 28 measures and you know, some complicated statistical issues with these measures as well. So I'm very much looking forward to hearing your expertise and feedback on the measures. But with that I do want to make sure everybody introduces themselves. And then I know Mitra actually has a few slides she'll go through.

MS. THEBERGE: Good morning everyone. I'm Suzanne Theberge. I'm a Senior Project Manager here at NQF. And I'm happy to meet to you all.

MS. ALLEN: Hi everyone. I'm Nadine Allen, Project Analyst. We worked on our previous work for phase one. And now I'm glad to be a part of phase two as well. I'm also working on the home and community based services project and the child Medicaid project.

MS. GHAZINOUR: Good morning everyone. This is Mitra Ghazinour. And I've been with NQF almost four years and supporting different committees, including the Measure Applications Partnership, MAP, Post-Acute Care/Long-Term Care Workgroup, and also supporting the new work on rural health. And I'm so happy that I'm also involved in this work, Person- and Family-Centered Care.

So I guess I'm just going to start with some introductory slides. We just wanted to go over quickly talking about Person- and Family-Centered Care Portfolio. And also discuss why functional status measures are under review for the person in this project. So, I currently -- the Person- and Family-Centered Care Portfolio includes 56 endorsed measures and measure sets. And during phase one the Committee reviewed 11 measures. 12 measures were submitted. One was withdrawn.

And the Committee reviewed 11 measures and 10 of which were recommended for endorsement.

And my understanding is that the measures have been reviewed by the Board and they are going to meet after this process. And so for this phase we have functional status measures to review. We have 21 endorsed functional status measures. And we have 7 additional new measures that were submitted during phase two. And we also have other measurement domains such as symptoms, symptom burden and other miscellaneous measurement domains.

So, why functional status is considered a measurement domain under person- and family-centered care? As you might be familiar with the work of Measure Applications Partnership, this is a multi-stakeholder group that is convened by NQF to provide recommendations on selection of measures for federal programs. And also to provide crosscutting recommendations, such as alignment, across federal programs, public programs and private programs.

So last year MAP convened three task

forces to identify families of measures for three NQF priorities of affordability, population health and person- and family-centered care. A family of measures are a set of related and available measures that address either high impact conditions or NQF priorities.

So the Person- and Family-Centered Care Task Force identified high priority areas -- five high priority areas. Did we lose -- okay, so I'm just to keep talking without the slides. So the Person- and Family-Centered Care Task Force identified five high priority areas for measuring person- and family-centered care. And quality of life was one of them.

So the task force emphasized that the importance of measures under the quality of life, including measures of behavioral, physical, social, emotional and spiritual well being. And also the importance of interventions designed to improve or maintain physical and cognitive functioning. And other sub-domains under quality of life included alleviation of symptoms and

symptom burden. And minimization of treatment burden on patients' families and caregivers.

So also another committee convened by NQF, which they defined patient-reported outcomes and also identified domains addressing patient-reported outcomes. So the committee defined patient-reported outcomes as any report of the status of a patient's health condition that comes directly from the patient without interpretation of the patient's provider or anyone else.

And the four domains included health related quality of life, including functional status, symptoms and symptom burden, experience with health care and their behaviors. So the next slide demonstrates the distinctions in terminology used to describe patients before the measurement.

The first one, as I referred to earlier, the concept of any report of a status of a patient health outcomes or health status that comes directly from the patient without interpretation of the providers. And the symptom

a patient might report that they have -- they suffer from depression.

And then PROM, which refers to instrument, scale or single-item measure used to assess the pro-concepts as perceived by the patient and directly reported by the patient. And an example is PHQ-9, a standardized tool to assess depression.

And then we have PRO-PM, which means pro-based performance measure that is based on PROM data and aggregated for an accountable healthcare entity. And an example includes percentage of patients in an accountable care organization whose depression score, as measured by the tool PHQ-9, has improved.

So the next few slides include all the 28 measures that the Committee is going to review today. I'm not going to list all the measures, name all the measures. However, these slides they show the display -- they display the breakdown of the functional status measures by setting. So the current slide demonstrates that

we have seven measures that address ambulatory/multiple setting rehabilitation. Such as a skilled nursing facility, outpatient rehabilitation.

And the next slide, it shows five measures that are applicable to home health and three measures that are specified for nursing homes and skilled nursing facilities. The next slide shows that we have seven measures that are specified for use inpatient -- we have inpatient facilities.

And lastly, we have two measures that address long-term care hospitals. And four measures that address outpatient settings. And also, I would like to go over some key points regarding functional status performance measures. So, surveys, instruments and tools are a method to collect data and not a measure by itself. And NQF endorses performance measures for accountable healthcare entities, not a survey tool or instrument alone. And a performance measure aggregates the data for the patient served by

each healthcare entity.

NQF endorsed performance measures are intended for use in both performance improvement and accountability applications, such as public reporting and pay for performance. And functional status, as I mentioned earlier, is considered a domain of person- and family-centered care. And we have a mixture of process outcome measures and patient-reported outcome measures for review in this phase.

There are some additional and key points including that measures can be based on single or multiple items, questions from surveys or instruments. And for outcome measures there's an exception to providing a summary of systematic review and grading of a body of evidence. And developers are asked to provide a rationale that at least one healthcare structure, process, intervention or service affects the patient experience being measured.

There are some key points specific to patient-reported outcome performance measures.

They are required to be tested at both levels of patient level data and score and the performance score for the healthcare entity. And lastly, PRO-PM developers are asked to provide evidence that the target population values the measure and finds it meaningful. And the last slide we just have included the list of patient-reported outcomes, performance measures that are under review in phase two. So before handing it over to Suzanne, I would like to know if there are any questions?

MEMBER MONROE: Are we -- when we look at these measures, we look at them as individual measures, not as groups of measures. I'll just use what's on the screen as an example for me. Should I be wondering is there enough difference between each of these to justify an individual measure? Or am I only looking at the measure and its properties and its value, et cetera? So that's just a question that I have.

MS. SAMPSEL: So, we will have that issue. And I think the important thing to

remember is we are considering these measures as individual measures. And then once we make a recommendation to move them forward or to endorse them or not, we would move into discussions about related and competing. So you do evaluate on the merits of the measure first.

MEMBER MONROE: Let's say we take 0422 and move it forward. And then we talk about related and competing measures.

MS. SAMPSEL: Yes.

MEMBER MONROE: Do we have the option at that point of then reconsidering whether we want to move 0422 forward because of the related and competing discussion?

MS. SAMPSEL: Yes.

MEMBER MONROE: So we do have a two-phase opportunity to look at these measures?

MS. SAMPSEL: Exactly. And in the related and competing discussions what will happen is we'll have a discussion, you know, is there an opportunity for harmonization? Or is it a discussion regarding depending on where the

measures match up, if they are related and competing to one another, where we might choose a superior measure, or a best fit measure. And it would be one over the other. And the other would not be endorsed. Correct.

DR. BURSTIN: And just one comment to add to that. That's great, Sarah. The other thing to consider is when we think about related measures, meaning they need to be harmonized., they're similar enough.

But if the patient population is different, as for example these would be, across different groups, they would not be considered competing because the patient populations are different. And that's where you'd want to make sure they are at least harmonized and make sense that the same sort of structure and method applies to each.

MEMBER KAPLAN: Can I ask a question? If some other group, like say there's a surgical group who's reviewing some measures of you know, performance measures for the surgery. And this

measure would actually enhance or in combination with some of those other surgical type measures create a -- is it NQF's job to go back and see if there's a mutuality there that can be enhanced? Or are you project specific? Are you bound by project?

DR. BURSTIN: Yes, it's a great question Sherrie. It's complex certainly. What we try to do is just put what we think is the right set of measures in front of the right set of groups and experts and multi-stakeholder groups. But at the same time we will oftentimes ask other committees to take a look.

And when we present the portfolio of measures for example to the surgery committee, we would add measures from this Committee so they could see the full view. But we probably need to do a better job of the sort of matchmaking of really thinking about how measures from disparate groups come together as composites or set of measures.

CO-CHAIRPERSON PARTRIDGE: And

Sherrie, actually you're anticipating what we hope will be a little discussion of this issue tomorrow afternoon, and which I spent some time discussing with my colleagues at the partnership yesterday afternoon. It's, I think we all feel that there's an expertise building in this Committee, particularly around the PRO-PM measures. Yet there are obviously technical -- clinical aspects that aren't necessarily reflected on the membership of this Committee. So how we put those together sensibly is an interesting question.

MEMBER THOMAS: In preparing for this meeting I read a letter that MEDPAC had submitted to the Department of Health and Human Services on the deluge of quality measures that are building and are coming into CMS every year. And it just got me thinking that we've got you know, take mobility. Mobility measures in home health. Mobility measures in SNF. Mobility measures in IRF.

I'm wondering you know, it's mobility.

There ought to be some ability to have a common set of instruments so -- or an instrument that cuts across settings, so that you can maybe eliminate some of the duplication of that. That's probably a naive comment but because I know a lot of this goes into each setting and accommodating the particular patient population that's served by those settings. But it just strikes me that therein lies I think a lot of the duplication.

DR. BURSTIN: This is a huge issue for the MAP, the measures of patient partnership as we're reviewing the measures that come forward. At the same time though again, you know, was much as we can harmonize the approaches, unfortunately we still live in a space where the data sets available in some of those settings tend to be quite different still.

Some of that is evolving, as you'll see today for some of the measures coming forward. But the key thing is to at least make sure that, however, its mobility is measured in

one setting should not be different then some of those key concepts. That's what we really think about in terms of harmonization. So I think that's going to be really important role for this Committee.

MEMBER PARISI: Actually Peter stimulated a thought in my head. Frequently in my experience and my observation, some measures are evaluated based on the perspective of the practitioner. For example, the way a physical therapist may evaluate or do an assessment of a patient's mobility, versus an RN.

So at what point does that figure into the discussion? Because some of these measures may be appropriate for a therapist, but not appropriate for a nurse or a physician. And that hasn't come up. So I was just wondering that question?

DR. BURSTIN: I think it's something you'll deal with. I mean in some instances there are going to be examples where there are measures directly from the voice of the patient, true

PROs. There are some where it's reflected through a nurse or a physical therapist. I think those are important questions to ask of the developers. Is this appropriate, it would be part of the measure specifications, that this is a measure done by X type of provider. Or maybe not.

I mean it could be pretty open. I think for example, I mean a set of committee measurements, PHQ-9 measures that we've endorsed, that look at the change in depression scores, are administered in a clinical setting to patients. But it doesn't prescribe who that person is.

There may be some instances for example, a physical therapist doing this work and doing the assessment, or a home health nurse doing that assessment, where they're the logical operator. But I think it's a question -- it's a fair question to be talking about through the day.

CO-CHAIRMAN STILLE: And I just want to thank all of you that have expertise in this

contents base. Because unlike our last set of discussions where we all kind of knew what patient- and family-centered care you know, was, some of these aren't as intuitive to many of us as they would be otherwise. So we'll probably be calling on more of you that have content expertise in this particular -- one or more particular areas depending on which measure we're talking about, to give us some perspective on that.

MEMBER BRADLEY: Could you help distinguish the -- because this was mentioned in several of the measure information that was presented. The difference between looking at the measure from a quality standpoint and an accountability standpoint? Because many of these measures are being proposed for payment. And so I'm just curious if you could define the difference. Is it pay for performance or are there other considerations under the accountability and payment issue?

DR. BURSTIN: I'm smiling just because

this is such a major issue for us at the moment at NQF. It's good I'm smiling. So, at this point in time, endorsement implies the measures are appropriate for a wide range of potential applications. Ranging all the way from quality improvement through all the way towards payment or penalties. We don't make that distinction. And that's how you should operate today.

That being said, the MAP process does in fact then take these measures and look specifically about whether they are applicable for a given program through CMS. Which will largely -- this is about a penalty, this is about public affording, et cetera.

So there is a second lens that offers that. The question we've really been grappling with, and I'd love your thoughts about this as we talk about this through the next couple of days, is whether we should also be moving the whole sort of evaluation process to being more about endorsement for intended use.

Or offering some gradation in the

endorsement that says this is a measure that meets the highest grades of testing, evidence, et cetera. And should be used for a variety of applications. These may not be quite ready for prime time for those, but could be appropriate for others.

That's still something worth talking about. And we're actually going to be convening an extra panel on that shortly. But for the meantime, assume it's the broad set of accountability applications. But know full well that actually next week the MAP will actually be talking about some of these very measures and their applicability to specific programs and whether they are reasonable for those programs.

MS. SAMPSEL: First of all, I wanted to kind of provide a little bit of context for some of the slides that Mitra presented. And there's a difference in this phase of work, which was the purpose of that slide, in that in the first phase of work, if you remember, we did a lot of patient-reported outcome process measures.

So we were looking for the item level of reliability and validity testing as well as the measure level reliability testing.

In this phase, it's only a few of the measures that are the PRO-PM. So we are looking at different levels of testing. So Suzanne is going to start talking and go through the measure evaluation guidance so we can start jumping into the measures.

But I do want to bring out, and we'll be bringing your attention back, when we are looking at PRO-PM and the level of testing that you're looking for based on the guidance that NQF issued last year. And there are some differences based on if it's a process, an outcome or a PRO-PM. And we have a mix of all measures.

So, you know, I wanted to make sure that folks understood that distinction because we're not looking all the time for the same things based on the different kinds of measures. So we'll bring that to your attention as we go through.

MS. THEBERGE: Okay. So I'll just talk real quickly about some process issues. We do have a quorum requirement of 75 percent of the Committee. We have 19 Committee members here, so that means we need 15 of you voting on any one measure to achieve quorum. So we do ask that, if at all possible, you not leave while we're in the middle of voting on a measure. If you need to step out, just try to do that during the discussion piece, because if our quorum numbers change during the votes it just causes some complications.

And also just to move something forward it needs a greater than 60 percent approval on any of the items. So that's at least 12 of you need to vote. And I think our new voting software will actually give us the percentages so we don't need to be doing the math as we go.

Next slide.

CO-CHAIRPERSON PARTRIDGE: Suzanne, I think we need to discuss a tiny bit about these

two other dots.

MS. THEBERGE: Oh yes, sorry. Sure. We have changed the process a bit, for those of you who have served on committees in the past. We now have something called "consensus not reached" because we had a lot of measures that were falling into the, like, 52 percent, 55 percent. And, you know, it's sort of like is that really consensus?

And so the three buckets that we have are pass/recommended, which is greater than 60 percent. Consensus not reached, which is 40 to 60 percent. And that includes both 40 and 60. And then does not pass/not recommended is less than 40 percent.

And so anything that's in that consensus not reached, "the gray zone" we call it, does continue to move forward. And we'll take it to comment and you will be asked for specific comments about that measure related to the consensus not reached status. And then the Committee will revote.

Of course, you have the opportunity to revote on any measure following comment. But you will definitely revote on those measures following the comment period. And just so you know, those consensus not reached, pass, does not pass, follows through at the NQF member voting and the CSAC as well.

So just process in terms of how we're going to move through everything today. The developers will briefly, two to three minutes, introduce their measures. And then our Chairs and Sarah will guide the discussion. We will have you comment on whether the measures do or do not meet the criteria. We ask that the lead discussant for each measure kind of run that piece of it.

If there are pre-meeting comments from the surveys that you all filled out, please refer to those during your introduction to the measure. And then we'll have you all vote on each of the criteria.

I think we already went over the

related competing thing. And then we can, of course, go over it again as it comes up.

Just some housekeeping things. Please make sure you turn your mic on when you would like to talk and then turn your mic off when you're done talking. I think we can only have two mics on at any one time. So that's why we need you to turn it off. And if you wish to make a comment, just turn your table tent up so that we know and our Co-Chairs will reach out to you.

And I'm going to turn it over to Sarah to talk about our criteria.

MS. SAMPSEL: Okay. So this is really meant to be a refresher course on the criteria. You know, as you'll recall, when we go through the votes and the discussions we'll ask you to, as the lead discussant, to introduce evidence and importance first and we'll vote.

We'll discuss scientific acceptability, which is the reliability, the validity, exclusions, et cetera. And we'll vote. And then we'll do feasibility and usability.

When Suzanne was mentioning the scoring thresholds, just as a reminder, if a measure does not pass importance or either of the scientific acceptability criteria for reliability and validity, that measure stops.

But then again, there may be a situation, and what we'll be doing is then asking you for some direct feedback to the developers regarding what it is it that stopped us so that we can give them direct and clear feedback on what they might do differently. Because there would be an opportunity for them to resubmit information prior to the end of public comment, which we had happen last time as well.

So, why are we concerned about evidence? Obviously it's the foundation for using as a quality indicator. And you know, it's a whole part of the validity testing and part of the validity requirements for is this a good measure and is this something we should be measuring in the first place?

For process and structure -- and,

again, we don't have any structure measure here, but we do want to make sure that each of these measures is something that healthcare units should be implementing. And I think we had a couple of questions earlier that were about that. I mean, are these measures important enough, is there enough evidence behind them that these are something that we should be recommending for use in the industry, whichever level of the industry they might be recommended?

And then when we're looking at the outcome measure, including those PR-PMs, which is the vast majority; I think that's 26 out of 28 of the measures that are either outcome or patient-recorded reported outcome. We want to make sure that whatever is being measured is something that the healthcare unit that's being measured, whether it's SNF, an inpatient rehab, long-term care, is something that they can influence and that there's evidence behind it. We want to see a tie between what is being measured to what can be done so that there could be improvement based

on what's being measured.

And we did, you know, focus a lot of our staff efforts in making sure that at least some information was provided on that. Was it the right information? That's up for us to discuss here.

Nadine, next slide. Why concerned about reliability and validity? Again, I think some of this has been discussed. That these measures are used in accountability applications such as public reporting and pay-for-performance.

We are going to have some examples. As Mitra mentioned earlier, some of these measures have already been considered by the Measures Application Partnership. And they're actually holding decisions on their measures on if they want to move them to final rulemaking based on the work of this Committee and recommendations for endorsement.

So, you know, these are important. And important discussions. And do want to make sure that we're moving forward reliable and valid

measures.

And then, you know, I think the rest of this slide is really just kind of issues that we'll talk about when we go through scientific soundness and scientific acceptability. But if we're moving a measure forward, we want to make sure the performance scores, you know, can be used to make conclusions. Because either the industry is using these measures to make quality improvement programs or to produce and move forward on quality improvement efforts. Or they are being used for pay-for-performance. Many of these measures are in some of the CMS Compare programs.

You know, I don't want to go through all of these notes on reliability and validity because of time at this point. We have created a cheat sheet on reliability and validity that we'll make copies of during lunchtime just to make sure you have reference. Because I know last time we had some questions, what is this Chronbach's analysis thing that you guys are

talking about? We have a new introduction this time of Rasch analysis, which is the first time folks have probably seen that. And so we do want to make sure everybody has the tools at their hand.

But we'll be asking the developers to comment exactly why they used a certain kind of testing as well so that we have that out. We'll be able to pull it up on the screen. It's actually on your SharePoint site if you want to pull it up. But at the same time, we've kind of spelled out and used some of the RAND tools because they do a really nice job of explaining this type of testing as well.

So I think we want to go into the measures at this point.

MEMBER KAPLAN: Sarah, sorry, I don't want to slow us down either because I know we're behind. But can I ask for some clarification about when a measure, for example, has been out there and it's up for reconsideration and it's now being moved into a performance category,

performance measure, is the guidance about -- because the guidance for an outcome measure may not be that you have to have empirical support of a link between process and outcome, or at least for the units being compared, but you can do a conceptual model or something.

If the measure is up for reconsideration and it's now moving towards a performance assessment, is there a requirement that now you must show, that for the unit being compared, you have to have some empirical support for that?

DR. BURSTIN: No, we actually don't. And it's been an interesting issue over time. So we have actually allowed outcome measures to move forward with simply a rationale for how they relate to the process measures, fully knowing that, in some instances, central line-associated blood stream infection probably being the best example, the outcome measure went out before a lot of the interventions that showed how you could reduce this.

I think there was a hesitancy to require that you have process improvements in hand before an outcome could move forward, recognizing the outcome at times can be the forcing function for its ensuring that some of those process improvements are discovered.

But it's been a contentious issue. And one we'll probably revisit many times over the coming years.

CO-CHAIRMAN STILLE: Okay. Sounds good. Let's dive in. We'll start with Measure 0423, Functional Status Change for Patients with Hip Impairments. And the FOTO folks will give a brief talk. And then, I believe, Sherrie, are you going to be the primary discussant, or Katherine? Sherrie, okay.

MR. JOHNSTON: Okay, great. Good morning, Madam and Mr. Co-Chairman, and members of the Committee. We thank you for the opportunity to present to you these seven measures that we are submitting today, Numbers 0422 through 0428.

I want to provide a little bit of history of how these measures were developed and how they're being used. And I'll first identify that FOTO started collecting data in 1995 by a consortium of large multi-state, publically-held healthcare providers of rehab. And they presented the data to the industry. And the industry said, that's great, but we don't like the data being held by a provider. So, resulting of that reaction, the data depository was put into the control of an independent entity of providers, which it has been since 1998.

Currently a number of patients starting an episode using these measures in the last 12 months was 1.23 million surveys or intake surveys. That's the volume of our survey process. This data is coming from over 15,000 clinicians, PT/OT and some speech, practicing in over 3,000 outpatient facilities in each of the states in the United States.

The survey platform is also being used by providers in the second largest HMO in the

State of Israel. And beginning this month, the Canadian Physiotherapy Association will begin to subsidize their member providers in the use of these measures to collect a standard data set.

The patient-report surveys are presented in seven languages. Of course, English, Spanish and French; Hebrew, Arabic and Russian.

The measures have evolved over the years. In 1995 we started with four legacy measures. The Oswestry for the lumbar spine, one for the knee, the Neck Disability Index for the neck, and the SF-36 for general health management.

In the last 20 years, to gain testing procession and efficiency in the clinic, we've added other anatomic-related patient-report measures. And to reduce the testing burden, we've gone to item response testing and processes and computer-assisted test technology to be able to successfully integrate the patient survey process in the clinical process.

The science of these measures was developed under the guidance of the late Dr. Dennis Hart, who served as the Director of Research and Development at FOTO since we started. FOTO's adherence to rigorous scientific methodology and psychometrics has led to the FOTO data being used at 89 refereed scientific publications.

FOTO may be the largest outcome database, measured externally, of outpatient rehabilitation. We currently have data on over 9.3 million patient surveys dating back to 1998.

We'd like to introduce the measures as a group, or at least the introduction as a group. The measures you're reviewing today are patient-reported outcome performance measures which use as their basis one of FOTO's patient-reported outcome measures.

The back history and research are very similar of reach of these seven patient-reported outcome performance measures. Beginning with 0422 for the knee and numerically advancing to

0428 general orthopaedic impairments.

Our performance measures are risk-adjusted and used at the patient and the clinician and the clinic level to assess functional level of a patient and the change in that functional level during patient care.

FOTO measures first received NQF approval in 2008 as a time-limited approval. And we received full endorsement in 2011. The PQRS has accepted FOTO's seven patient-reported outcome measures. And FOTO is qualified as a PQRS data registry.

Our application today includes some revisions, primarily moving it from what we think is a process measure to an outcome measure. And lowering the age from 18 to 14. And there's a few other small changes outlined in there.

Because I am also a physical therapist but have no expertise in research or the science of this, I have brought with me a presentation panel. Jerry Connolly back here, our consultant for public policy. Dr. Mark Werneke, who is a

clinician at Central State Medical Center in New Jersey and also a researcher. And Dr. Daniel Deutscher, Director of Research from Maccabi Health System in Israel. And hopefully on the phone is Dr. Linda Resnik, who has been the leader of this distinguished team and this challenging effort.

And, finally, FOTO thanks the NQF support staff, Mitra and Nadine, for their wonderful patience and their guidance and cooperation during this application process. In addition, they have helped us add additional measures to make our application more complete and to improve our measure analysis. Thank you very much.

CO-CHAIRMAN STILLE: Okay. Sherrie?

MS. KAPLAN: Hi, thank you for that introduction. I had some confusion starting off in trying to describe the measure. The term residuals to somebody like me -- and I'm statistically trained -- so a residual to me means unexplained variance. And what I think you

mean is that it's a changed score adjusted for certain characteristics of the patient. Is that correct? It's not a residual, right?

MR. JOHNSTON: Could I confirm that Linda is on the phone, Dr. Resnik?

DR. RESNIK: Yes, I'm on the phone. Do you want me to take that? Or, Dr. Deutscher, would you like to answer? Can you hear me?

MR. JOHNSTON: No, you couldn't hear me because I turned my button off. I'd like for you to determine whether you should answer or whether Dan or Mark should answer.

DR. RESNIK: Okay. Well, I'll take that question. After the risk adjustment process, what's remaining in the model is variation from the predicted value. So that is the residual score after modeling. That includes error and what we believe is the variance due to clinician and clinic characteristics. So that's what the residual is. So it's the risk-adjusted value after the modeling. Does that answer your question?

MEMBER KAPLAN: No. Now I'm more confused. Because you say that it's the change between the intake and the discharge value, so in -- help me out with the model here. What's the dependent variable? A dependent variable -- let me just sort of frame my confusion.

If the dependent variable is discharge, functional status at discharge, in the model do you include the baseline measure, the intake value, along with the other adjusters? How do you compute the -- or are you looking at the -- are you really looking at residuals, unexplained variation after adjustment? And what's in the model?

DR. RESNIK: Okay. The risk adjustment model is specified. It includes intake functional status as well as key characteristics that are specified in the model: gender, age, comorbidities and so on, acuity or onset. And the dependent variable is change. Change from intake to discharge. So that's what the model looks like.

MEMBER KAPLAN: Okay. So in the model, on the right-hand side of the model, is intake value? And on the left-hand side is the change score?

DR. RESNIK: Yes.

MEMBER KAPLAN: So why isn't that an over-specification of the model? Because you would not -- I'm sorry, I don't want to get into the details here, but it's important, I think, for the Committee to understand exactly what's being evaluated.

If you put the -- the thing is predicting itself if you've got the intake value on the right-hand side and you use the intake value to compute a change score on the left-hand side.

DR. RESNIK: No, the intake value is not on the right-hand side. The right-hand side, the dependent variable, is change. So it's the difference between discharge and intake.

MEMBER KAPLAN: But you then don't have the intake value on the other side of the

equation.

DR. RESNIK: Yes, we do. Because change is dependent on the baseline status of the patient. So patients who come in with a great deal of impairment may change a different amount than patients who come in with minimal impairment.

MEMBER KAPLAN: Okay. I think that could be a problem. But we're going to probably need some more discussion on that.

MEMBER BIERNER: Can I ask a question there? So are you saying that the same delta, the same change, or the same measured change, in a given person will vary depending on what their disability is or what their level of impairment was when they started?

DR. RESNIK: Yes.

MEMBER BIERNER: Okay. So can I compare, if I were looking at a physical therapy clinic, can I take the scores that are generated from this measure and compare it to another clinic without knowing much about their patient

population? Would I have to know something a priori about the kinds of patients they see? Say they're severely disabled versus an outpatient sports medicine clinic. Would that delta, that change, be the same in those two settings or not? Or would there be some modification based on where the patient was when they came in?

DR. RESNIK: Right. Because the intake score is in the model, that accounts for the functional status of the patients within the clinic at intake. So that's why we have the risk adjustment model because we know that different clinics serve different populations.

MEMBER BIERNER: Okay. But it is possible to compare -- I just want to understand that I can compare apples to apples that you're --

DR. RESNIK: Yes. That's why we have --

MEMBER BIERNER: I understand what Rasch analysis is and what you're trying to do, as I understand it, is to make sure that we are

comparing apples to apples, that we're comparing severely disabled to severely disabled and not sports medicine clinics to hospital-based PT, for example.

DR. RESNIK: Right. For each individual patient, each individual patient has a risk-adjusted score. So, in other words, for each individual patient we can, based on their intake status, age, gender, symptom acuity, surgical history, comorbid conditions, fear avoidance beliefs, payer, we predict their outcome. And then we understand the difference between their actual outcome and what's predicted.

MEMBER BIERNER: Okay, but --

DR. RESNIK: And that's for each individual patient. And then by clinic, we aggregate the risk-adjusted or residual scores by clinic so that we are comparing the predictions for individual patients within clinics. And so we're taking into account the patient characteristics within each clinic.

MEMBER BIERNER: Okay. So that means, though, that, in addition to the actual questionnaire or instrument that we're being shown, there's a lot of other demographic or other information you're collecting that's not shown in this instrument.

DR. RESNIK: That's right. The patient inquiry tool that's used by FOTO has a key component of the survey where we assess information on age, gender, onset of the condition, number of surgeries for the condition. We have a list of comorbid conditions known to be associated with physical function. We have the type of payer, we have other surveys, like fear avoidance beliefs.

Those are all accounted for in the risk adjustment process. So, yes, there are a suite of other survey items that are added into the model. Those are not shown.

MEMBER BIERNER: Okay. So I think that's important for this Committee to understand. Because I'm understanding it now,

but it wasn't immediately obvious from reading the material submitted. I mean, that's why -- this is something that -- they're buying into your product because in order to collect all that other information and use your large database to analyze against prior history, that's what you're doing. It's not just this one instrument.

DR. RESNIK: Yes.

MEMBER KAPLAN: So, let me just kind of review the risk adjustment while we're on that. And then I want to kind of move us to a couple of other concerns I had about this measure that I need some clarification on.

One is that the risk adjustment scores -- I understand that the risk adjustment modeling you did was for age, gender, symptom acuity, surgical history, number of functional comorbid conditions, payer and level of fear avoidance beliefs of physical activities. And that's it, right? There are no other things that you're measuring in that risk adjustment that we need to understand?

DR. RESNIK: I think that's it, yes.

MEMBER KAPLAN: Okay. So, but does the etiology of hip impairment, is that included? For example, you say surgical history. Is the etiology of the hip impairment, is it a hip fracture, or is it just osteo-whatever, arthritis? Or is the hip impairment etiology included or is it just surgical history? So, does it matter? And I'm thinking about things like, well, if all the care prior to the time they hit your intake observation point could be the lion's share of the predictor of what the recovery trajectory looks like, then we're missing key information.

DR. RESNIK: Right. There is no diagnostic information taken into account in this risk adjustment model. And I think that the work that's been done in this area demonstrates that -- and I do believe etiology may be important. However, etiology is reflected in the intake functional to a great extent. And so we're able to predict a fair amount of the variation in

patient outcomes just with these characteristics, without diagnosis or etiology.

MEMBER KAPLAN: Thank you. The R-squared values are, if I recall, .37, .35, something like that. But --

(Simultaneous speaking.)

MEMBER KAPLAN: But if the intake measure is in there, then we would like to see the additional information accounted for by all the other variables taken as a group. Because if the thing is mostly predicting -- I mean, the best predictor for most functional status measures of future function is prior function. If that's accounting for most of the variability, the residual variance may largely be attributable to error.

DR. RESNIK: We could speculate that. And we don't present here, but we have looked at hierarchical models where we do see that a certain proportion is attributable to the clinic and the clinician level. And I have done that in my own research, yes.

MEMBER KAPLAN: Okay. That is something that I also wanted to get some thoughts from the measures' developer on. Because only patient-level data and references are provided linking the performance measure to interpretable variability.

And there's an assertion on page 22, and again on page 23, paragraph four, that the use of the measure for performance improvement at the provider level sort of "makes sense." And although you give some lines and stuff at the end at the clinic level, I have a lot of issues about whether that's enough justification given some of the analyses you've done.

So let me just kind of walk you through what I found. And then other Committee members can chime in behind.

With respect to things like missing data, you've got a 50 percent attrition rate between intake and discharge. And absent links between treatment intensity, because you do have the number of visits, the patient --

DR. RESNIK: We don't have the number of visits necessarily.

MEMBER KAPLAN: Okay, but you --

DR. RESNIK: Not in the model.

MEMBER KAPLAN: So help me out with who is in the sample. Because you make some statements about physicians have to have at least ten patients per -- right -- per physician.

DR. RESNIK: Yes. For clinician, yeah.

MEMBER KAPLAN: Okay. For clinician, sorry. And then at the clinic level you have to have at least -- for clinics smaller than five clinicians, they have to have ten patients per visit. But then for clinics larger than five clinicians, there was another sample size estimate. Forty completed episodes --

DR. RESNIK: Yes. Let me clarify. In the risk adjustment modeling that specifies the entire model and what the coefficients of the model will be, we use all patients who have complete discharge, intake and discharge scores.

But then to calculate the performance measures, we have thresholds for participation. But then we have the rules about clinicians and the number of patients they have to have each, and clinics and the number of patients, so that we have a more stable estimate for the performance measure. But our entire risk model uses all patients.

MEMBER KAPLAN: Okay. The specification on that, and the sample to whom it applies, I think we need more information about. At least I would feel more sanguine about it if we had more information, because instability of measurement, the fewer observations you have to sample from, obviously, the more measurement error you're going to make.

So people who come more often are more likely to get sampled and so on. And they're more likely to get care, which would help interpret the link between -- at least intensity -- between the process and outcome issues.

So I got real confused about what we

were kind of -- how we were sampling things. And then whether or not you did hierarchical models, how are the splines in Figure 2B, 52A, on page 17 of the attachment? What statistical approach -- did you use generalized estimation equations? Hierarchical linear -- how did you generate those lines?

DR. RESNIK: I'm sorry that I'm unable to see the slides that you're referring to. Daniel, if you're there, I believe these are your figures that you generated. And I'm fairly certain that these are models that are not from hierarchical models. These are the results of another analysis. Daniel, can you take that?

DR. DEUTSCHER: Yes. You were asking about hierarchical models. But we're not presenting any here. So we did not use hierarchical models for these applications.

MEMBER KAPLAN: But you have a nested design. You've got patients with -- the way I understand it, you've got patients within clinician and clinicians within clinic. So why

would you not use a hierarchical approach?

DR. DEUTSCHER: Well, since the risk adjustment model is used to calculate the prediction -- the predicted score -- I think you're absolutely correct on a research basis. But for applications on a day-to-day basis, if you want to provide a risk-adjusted change score, a risk-adjusted discharge score, that would be difficult for the clinicians using the software to take into account the nested models.

So that's something we thought of looking into. But we haven't done that yet in order to move that to a practical application in a routine clinical environment.

MEMBER KAPLAN: Yeah. My concern is the misinterpretation possibilities of interpreting small, very, very, very small numbers at the individual level certainly is going to get very noisy very quickly and be a real problem, if that's the intended use.

And my understanding was you had to have a year space between intake and discharge,

right? Does that vary?

DR. DEUTSCHER: Well, yes. First of all, I think it needs to be clarified that when we run the model, in order to achieve, to get those coefficients used for the prediction, we use all patients in the database that have scores at intake and discharge. So that's based on a very large sample size.

But then only for reporting purposes, because of what you've just said, we do not provide reports at the clinician or clinic level if they do not pass a certain threshold, because of this worry. We worried about very small sample sizes giving very unstable estimates.

MEMBER KAPLAN: And that's exactly what hierarchical linear modeling is designed to do, is to take into account the differences and standard error measurement that you would get when you have floating sample sizes. So I'm kind of surprised that you didn't use that technique.

Let me just follow up. The standard error of measurement at the clinician level may

not allow you -- the splines you projected look like you can't really discriminate except for the very, very ends of the distribution. And so you would mostly say that most things don't differ. They're not -- and the use for a performance measure means the thing really has to tell who's what. And certainly not at the individual level. So how do you envision the quartiles, deciles, five percent high and low?

DR. RESNIK: I guess I would disagree with your comment and say that we do demonstrate that we can differentiate three groups who do not have confidence intervals that overlap. And those we would say are people with average, or as predicted, outcomes of their patients. And then clinics with better than predicted and then those with below.

So I think we see three groups whose confidence intervals around this estimate of their patient outcomes do not overlap.

MEMBER KAPLAN: But what -- first of all, I'm not sure I can see that from what you've

generated. And second, how would you use that information? I mean, if those confidence intervals are great, but if they give you -- you would chop it into three groups and how? What distributional properties would be true over different observation points? Would they float?

Would you have -- you know, I'm now getting more confused than I was at the start about exactly what it is you're asking us to approve.

CO-CHAIRMAN STILLE: I'm going to need to pull us back a little bit. This is great discussion about validity. And actually let's bookmark that, because when we talk about validity all of this is going to be relevant, plus some of the other measures.

We need to kind of talk about importance in priority and stuff, because we need to vote on that, I believe.

CO-CHAIRPERSON PARTRIDGE: We have other discussants and we all have looked at -- I would -- since I'm one of the reviewers of other

of these measures, I'm very glad Sherrie went first. And I am because I think -- I don't know if I speak for others -- but I have question marks all over my particular review. And I'm getting nods around the table from some of the rest of us.

I think, for the perspective of our developers, we're really struggling with this one, I think, to understand pretty clearly what the numerator is, what the denominator is and how you calculate each one of them. And, Len?

MEMBER PARISI: It would also be helpful to understand the overlap on the methodologies across all the measures that are related so that we don't have to repeat the discussion.

CO-CHAIRMAN STILLE: Right. I'm guessing there's a lot.

CO-CHAIRPERSON PARTRIDGE: Yes. And for example, I think I've picked up what ODQ was in your discussion. I think you mean the Oswestry Disability Index. Am I right? That you

talked of using different intake tools depending upon what we're looking at. You use a different intake tool perhaps for the lumbar from what you use for knee. Am I right?

DR. DEUTSCHER: Well, the tools are basically different.

CO-CHAIRPERSON PARTRIDGE: They're standard tools. I understand that.

DR. DEUTSCHER: They're a combination -- sometimes combinations of standard tools.

CO-CHAIRPERSON PARTRIDGE: Okay.

DR. DEUTSCHER: For instance, for the lumbar, the lumbar computerized adaptive testing measure was created from the lower back pain functional scale, included also some items from the SF-36. They were all combined using an item response theory methodology, a Rasch analysis, to see unidimentionality and things of that sort.

CO-CHAIRPERSON PARTRIDGE: The reason, I think ,in some of the other measures we're reviewing, we some of these standard tools coming up again. Not used necessarily in the same way

you have used them. So it's going to be useful for us to sort it out.

DR. RESNIK I think we need to make it clear that the FOTO measures, while they might have had items that originated in some of the other tools, and in the lumbar application, we did present a comparison between the FOTO PROM and the Oswestry.

The FOTO measures are unique in that now they have gone beyond the original items to be computer adaptive tests or short forms based on the items test, the computer adaptive tests. And there's been, for most of the measures, extensive publications on the psychometrics of the development of the FOTO patient-reported outcome measures. So they're not the same as the Oswestry.

DR. DEUTSCHER: Could I add just a clarification regarding this point?

CO-CHAIRMAN STILLE: Go ahead. Yes.

DR. DEUTSCHER: It's important to understand that, as Linda Resnik has just

described, the measure itself is combined from several measures. Not all of the items were combined. But all of the data that's presented here that's been collected are data that were collected using the FOTO combined measure. Not separate measures. Not the Oswestry and then some other measures.

CO-CHAIRPERSON PARTRIDGE: Yes. Yes, I understand.

CO-CHAIRMAN STILLE: Yeah, let's have a brief discussion of the importance stuff and then we can vote on that. And then we can go back to the validity things.

MEMBER SALIBA: Thank you. I have a question. Can you clarify whether or not is this in the public domain? Is this quality measure in the public domain or is it copyrighted? I wasn't clear.

DR. RESNIK: Yes. As the application shows, each of the measures has short forms that are in the public domain. We have links to the FOTO website where the measures are available in

the public domain. And the risk models that FOTO uses are also available to anyone in the public domain.

MEMBER BIERNER: I wanted to ask a question about the small number of workers' compensation patients, it looks live five percent was approximately what you had. Is that correct?

DR. RESNIK: I think it varies by the PROM.

MEMBER BIERNER: And the one under discussion is the hip one is the one I was looking at. So I think it's five percent. But I'm not trying to talk about all of them at the same time. But the one under discussion is the hip one.

DR. RESNIK: Yes. In Table 1.6D, there's been so many analysis with the FOTO data that we have different samples for different of the analyses that has taken place over the last few years. So, in the first table, we study three percent with workers' comp.

MEMBER BIERNER: My point is that the

use of this measure in the future, in a setting involving workers' compensation, might not be accurate because of the small number of workers' compensation patients that you have in our data set.

DR. RESNIK: I think we should probably look to a different table for -- because that particular table, the first table, was a test/retest reliability sample, which was very small. And I'll try to find another table to confirm whether that is the case or whether it is still the same small --

CO-CHAIRPERSON PARTRIDGE: I think before we move off importance entirely, I would like to understand myself -- and my fellow reviewers, if you all get it, stop me. What's the gap? Let's start with the first measure, 0423. The three elements under importance include link to procedure, performance and prevalence.

DR. RESNIK: Right. We understood that we did not present this clearly in the

application. And we have submitted some supplemental materials to address the gaps and disparities. So you'll see there is handouts. There's a table called Disparities, data tables, and I don't know what their handouts look like. But this shows differences in outcome between age group, gender and payer type for each of the measures. So if we're talking about the hip --

CO-CHAIRPERSON PARTRIDGE: So if I am trying to decide whether I want to use this clinic or another clinic post-surgery, and I go to your website, what went into saying that this clinic is an A-1 performer as opposed to others? I'm sorry, I'm getting into the data again. I take it back.

Sherrie, are you comfortable with their -- with the state -- she's shaking her head, around GAP?

MEMBER KAPLAN: No. And it's because I couldn't understand how the scores were constructed. And I still am confused about how -- I'm not sure what we're trying to -- what we're

being asked to endorse.

And I didn't hear what the mode of administration did to the scores either. Because there's paper and pencil in here. There's IRT and CAT generated. Are we endorsing hip CAT? Are we endorsing the paper and pencil version? How are those related? What's mode of administration doing to all of this?

I don't have enough information to know what I'm being asked to vote on.

MEMBER NEUWIRTH: Can I ask a question as well? I was trying to find the actual instruments that were used. The hip CAT. And I couldn't find the link in here. I found it for the knee.

MEMBER BIERNER: I found it on the FOTO website.

MEMBER NEUWIRTH: So you had to go to the FOTO website. Okay, so --

MEMBER BIERNER: Yes. I did find the instrument though.

MEMBER NEUWIRTH: Okay.

MEMBER BIERNER: The paper one.

MEMBER NEUWIRTH: The paper one. And is it --

DR. RESNIK: The paper form -- I'm sorry, the paper forms are short form versions of the CAT measures. They predict about 96 to 97 percent of the variance of the full measures.

So we believe that they are equivalent or roughly, very close to equivalent. So in terms of the mode of administration, there's not a lot of any bias introduced or minimal bias.

MR. JOHNSTON: Well, and also on the website is a link to the actual survey, the CAT survey. And we believe that we've placed it in the public domain as well.

CO-CHAIRMAN STILLE: Yes, I found it on the FOTO website while we were talking. Is there a sense of the Committee that we're ready to vote on any of the initial measures such as Importance?

MEMBER BEVANS: Can I make one comment as a discussant? This is an issue that I think

does fall under importance. Probably though not just related to this version of the FOTO because just to bring up the more general point of whether functional status measures should include attributions to specific body parts.

So you know, not just this measure but the suite of them. The advantage of course being that it has a likelihood of greatly enhancing measures for, you know, the treatment specificity perhaps.

But of course that assumes that the body part approach -- assumes that a change in an individual's functional status can be attributed to that specific -- the function of that body part and that that is well understood by the patient. It also limits the degree to which comparisons can be made across clinics that are treating, you know, people for different sort of body part injuries.

I wanted to bring that up here. I realize it's something that applies to you know, the entire suite of instruments, not just that.

But I think it has some pretty important implications for the Importance of the measure.

DR. RESNIK: In my case I would argue that functional status may be a recent construct and we all can be comparable. However, conditions, say of the wrist and hand, affect functional status quite differently than conditions of the foot and ankle.

And because we want to be brief and be able to measure things in an efficient manner, we choose different items to get at that construct for people with hand impairments as compared to foot impairments. And for people with back impairments as compared to foot impairments.

So I think it's the selection of items to get the most efficient and accurate assessment of the aspects of function that are affected by impairments in those body regions. And that's why we have the different body part specific CATs.

MEMBER BEVANS: I get that rationale I think though that, you know, a lower extremity

mobility type construct or an upper extremity could also work.

CO-CHAIRMAN STILLE: And it's interesting because when you look at the --

MEMBER BIERNER: Well, there are already measures. There's a lower extremity functional scale, there's a DASH. There are other measures that cover like the upper extremity.

There's the Womack for the hip that I was going to ask about on this one. But so there are already other measures. But I'm supportive of the fact that you have to have different specific questions and different aspects of functional impairment that are specific to body parts because there are a lot of differences in a hip patient versus a foot and ankle injury patient or a hand or shoulder.

These all have significant differences and the kinds of impairments they -- or problems of daily living that they have. But there are some.

So I'd like to know how does this compare to the Womack, which has been around a long time for the hip. Did you all -- has there been any head to head testing? Or are you all just started with when you first started this?

DR. RESNIK: I don't think that we have ever directly compared to the Womack. And the Womack is not also -- it's not an NQF endorsed measure.

I mean there are many, many functional scales. And we haven't compared to all of them. That would involve some time collection of data and different aspects.

MEMBER BIERNER: Well, yes. But the Womack has been around 30 years. And it's one of the most well known for arthritis of the hip and knee. The hip and their knee measures have been around for you know, decades, and are very well published.

CO-CHAIRMAN STILLE: Ann, did you have a question or did you just put your thing down? No. There's three more up, Len, Liz, Sherrie?

MEMBER KAPLAN: I just have a feasibility question and it's a quick one. Because that's why I asked the question about how much these other risk adjuster variables are important in explaining differences in these measures? Because you've got a 15-point improvement on a zero to 100 scale over a year's period.

Because nobody that I know collects routinely level of fear, avoidance beliefs of physical activities. So does that tie this to this -- to FOTO in a way that makes it unusable by a larger group of folks because of your risk adjustment model?

DR. RESNIK: The reason that fear avoidance is in there is because we have found that it is predictive of outcome. And that patients who have higher levels of fear avoidance do not do as well in therapy.

And so to equalize clinics that may see that type of patient, we do feel that it's important to adjust for that. The measure that's

used for that is not lengthy and is available.

MEMBER SALIBA: So fear avoidance is being used as an independent variable?

DR. RESNIK: Yes.

MEMBER SALIBA: Okay.

DR. WERNEKE: The amount of time -- go ahead Linda.

DR. RESNIK: Sorry.

DR. WERNEKE: The amount of time it takes to collect this information using the CAT is about one to two minutes. And as a clinician I used to collect information with the Oswestry, et cetera, and that took six to eight minutes for the patient to complete it, for the clinician to record it and then to try to interpret it.

And we did head to head comparisons between the FOTO CAT and the Oswestry. They behaved similarly psychometrically, but the burden of using a similar tool was so reduced. It makes it so much more efficient.

And a matter of fact, the tool is so efficient I also collect biopsychosocial surveys

for all of my patients. And having that CAT efficiency is what allows me to do that.

So my specialty is in low back patients. And I collect multiple psychosocial factors along with the physical functioning scale. And I can do it very efficiently and the patients have not objected.

I had one other comment too about specific -- body part specific. This was really driven by customers and clinician input. Where they wanted to stop irrelevant items being asked to their patients. And the CAT enables us to do that.

And the clinicians were demanding a more efficient tool. And that led to the push to the development of the CAT.

MEMBER BEVANS: I'm glad you mentioned that. One of my questions was related to content validity. Unless I missed it, I didn't see any information about patient input or clinician input into the development of the actual items. And you know, verification that those are

important, meaningful constructs for people, so.

DR. WERNEKE: Yes, the clinician was involved. We canvassed not only patient managers as well the clinician input. And again, about relevancy and burden was very important.

In fact, when we asked patients what they thought of using the CAT, they were very happy first, oh wow, I get one question at a time. I don't have to see the whole ten questions on one form. And they liked the large font size.

The other positives were from the clinician. And again the manager was the efficiency. The only drawback that we heard from the managers was the cost. And then a fear of interfacing with the older population with the computer.

And that was resolved in about 2005 when we started recommended the penlight and touchscreen. And I've been using that for a long time now. I have no problems having my older patients connect with use of computer

administered surveys.

CO-CHAIRMAN STILLE: Liz and then Len and then we should probably vote. Start to vote.

MEMBER MORT: On the issue of body part specific measures, I'm all in favor of body part specific measures taking care of patients. They really care about the function related to that body part.

So I actually have a question in the more disaggregated area of why the body part is affected. And this relates to what Dr. Kaplan was saying. When I look at -- I was the shoulder. I had the shoulder one.

And when I saw the variability in the injury and disease types of shoulder problems that were included in the denominator. And being a clinician, I just can't believe that you can actually risk adjust that difference away.

And therefore, when you are looking at an individual patient I would say fine. But if you're trying to aggregate anything that's that heterogeneous into a measure about a clinician's

performance or clinic's performance, I think you're just getting much too much variability that cannot be controlled adequately from the risk adjustment model as specified.

So I had trouble from the evidence perspective with that issue. I wondered if the developers had anything to say about it. But I must say, I do love your items.

Yesterday about 30 percent of my patients had shoulder injuries. They all went to physical therapists. And these are the things that they can't do. Lift, comb their hair, hair dryers and that sort of thing.

So I think you're really onto a very important area of people's function. But I have questions about the measure.

DR. DEUTSCHER: Okay. The actual truth is that the biomedical model has difficulties explaining exactly when somebody comes in and says I have a shoulder pain, what exactly is the source of that pain. And we know that the validity of many of ICD-9 codes and we

saw a lot of those codes in the applications is questionable.

What we do know is that the selection is the selection of the patient. They come and they say my shoulder is my main problem. And we also do know from the IRT analysis and the Rasch analysis that those measures function as unidimensional as possible.

It's never a fully unidimensional measure. But the unidimensionality is maintained as far as we can assess it. So I don't know if it's even possible because there's always a variability that we won't be able to explain.

I don't know if it's even possible today with the knowledge we have today to say exactly for a specific patient in a reliable and valid way, is it your labrum that's affected. Is it you know, the tendon or not.

Many studies have shown that that's not really possible. But as you said, when the functions themselves that are assessed, they relate to actual problems the patient have, the

measure becomes unidimensional.

MEMBER MORT: Well, but just imagine a 21-year-old man who skis and breaks his clavicle. And versus an 84-year-old man or woman who has adhesive capsulitis that's related to degenerative joint disease and not you know, inactivity.

Two entirely different -- both have pain. And maybe both can't do hair dryer and things. But they're very, very different. So I guess the clinician in me and the measurement person in me just finds difficulty in combining all of them around the symptom.

DR. DEUTSCHER: I agree that they're very different and we do risk adjust for age. But there are other differences that we might not be able to risk adjust for. But the question is which functions are they trying to achieve?

And many of the times even if the source of the injury is different, the functional tasks that they're trying to achieve are similar. And this is what we're actually measuring. How

do they perceive? What's the difficulty level that they perceive regarding specific tasks?

CO-CHAIRPERSON STILLE: Len?

DR. RESNIK: And I guess I would also add that because we adjust for onset of the condition, we would know that the patient who had the clavicle fracture was you know, had an acute injury. And we would also adjust for comorbid conditions. So we would know that the older person had a, you know, had a condition of arthritis.

So there is more than just the intake function into account in the models. Although certainly it's not perfect. But diagnosis codes are fraught with error.

MEMBER MORT: I guess I was thinking more of a stratification or having, you know, separating the populations. If I were the 21-year-old with the ski injury, I would want to know does that physical therapy rehab facility take care of people like me who are otherwise athletic and just had an injury.

If I was the 85-year-old, I'd want to know does that group really care about the elderly and our function as we get into our senior years. So stratification might be another approach.

CO-CHAIRMAN STILLE: Len, thanks.

MEMBER PARISI: I'm actually on a related question. As it relates to the GAP. I know we touched on it. But from a quality perspective it's not clear to me the connection between collecting this information and how it drives improvement apart from individual clinicians driving that improvement with individual patients.

So I'm not seeing the connection under the GAP. So if you could help me with that that would be good. And not only from the measure that I reviewed, which is 0424, but also for all of them.

DR. RESNIK: We did present another supplement on clinician performance over time. There is a supplemental handout where we looked

at clinician performance for clinicians who had been stable subscribers in the FOTO database over time to see if participating in the system and getting the feedback on their performance actually changed their performance.

And what we can see in that slide is for clinicians who again had a minimum of ten patients a year and participated for all three years, we see that -- an improvement in performance over time. A greater proportion of those clinicians moved from a lower and average performance to high performance. And you can see that in the handout.

So we think that just having the information and feedback on your performance as it relates to what's expected and what your peers are doing, does drive performance of the clinician and clinic level.

CO-CHAIRMAN STILLE: Sherrie?

MEMBER KAPLAN: I was trying to say nothing more this whole rest of this discussion. But that particular question bothered me.

Because where -- what was the -- what were the interclass correlation coefficients for measure 0423 at the clinician level?

I know they're very high at the patient level. And at the patient level I real -- this doesn't bother me at all. When you start using it at the clinician and the clinic level, that's when I have concerns about using this as a performance measure.

I don't think we've got enough information about that issue. What were the ICCs at the clinician level?

DR. RESNIK: We have not calculated that yet. And we will do so and submit to you.

MEMBER KAPLAN: And the clinic level?

DR. RESNIK: Yes, we will present it at the clinician and the clinic level. We will calculate those. But we have not done so to date.

CO-CHAIRMAN STILLE: Okay. So this is actually a perfect segue. Because I was going to ask Sherrie and Katherine, what recommendations

they'd have for the measure developers to come back to us with things. If you had a checklist you'd like to maybe give them?

MEMBER KAPLAN: Well, that was number one. I would like to see what the ICCs are. Because with the interclass correlation coefficient is the different -- the between unit variation divided by the between minus within unit variation.

So if there's a strong clinician thumbprint and there's a lot they -- they tell the same story across all patients, then they're small within clinician variation. And if the measure then distinguishes lots between my colleague down the hall who does the same thing but way differently than I do who does the same thing, then that number will be fairly large.

And what you'd like to see for performance measures that are now being used at a different level, are the interclass correlation coefficients to make sure that there is enough evidence that these are distinguishing clinicians

from each other.

The second thing I'd like to see is some evidence of validity. But well, maybe that's not possible. Maybe my first checklist would be the interclass correlation coefficients. And then some link with either -- if you have intensity.

If you have some kind of visit information at all, the idea that clinician -- patients who get seen more frequently are doing better. And that physicians who see patients on a more frequent basis have higher scores or some -- some evidence of validity outside of -- at the clinician and the clinic level, not the patient level.

DR. RESNIK: At one point we did also submit another piece of supplementary information that showed some additional validity of the provider classification method. And it's called -- the handout was called the link to the provider classification.

And I apologize if you may not have

seen it. But we did it. We looked at our classification of low performers, average performers and high performers.

And then we looked at those clinics what percentage of patients had made improvement in their functional status that was greater the minimally important -- minimally clinically important difference. And we found as we expected that clinics that were high performers, a greater proportion of patients improved greater than a minimally clinically important difference.

So we presented that type of validity evidence at the clinic level by year.

MEMBER BEVANS: In addition to the reliability information at the aggregated level, I have two other requests. One a justification for the risk adjustment variables. I'm concerned about a couple of those. Specifically gender and payer.

Because I don't know truly what the evidence is with regard to potential gender differences. For example, in the speed or

magnitude of improvement. And I'm concerned that both gender and payer risk adjustment may actually mask some of the important disparities in the quality of care provided.

The other point I wanted to make is the instrument has been modified for use originally with 18 years of age plus, down to use with youth I think as young as 14. But there's no evidence provided and perhaps you have it. But it's not provided that the measures have been tested for understandability and appropriateness with adolescents. That's on my list.

MEMBER MONROE: Just quickly, I'd like to pick up on Deb's earlier question. I just want to understand, when she asked if it was in the public domain, I think the response was that there's a short form that's in the public domain.

How different is that from the full measure? And why the distinction?

MR. JOHNSTON: Well, the short form, it's composed of a portion of the items that were considered to be the most important items in the

full item bank for that particular measure.

MEMBER MONROE: Has that been tested similarly to the full set?

DR. DEUTSCHER: Yes, the scores from the short form are -- were calibrated using a cross -- kind of a crossover table to the original CAT.

MEMBER MONROE: I'll have to rely on my statistician colleagues to tell me if that's an appropriate answer.

MEMBER BIERNER: What it is really, there's like a data bank of questions. And when you take the computerized test they can choose multiple questions. The short form on paper has ten questions. I was able to pull it up.

And those -- as he's saying, those ten will give you an equivalent score statistically compared with their computerized test.

DR. RESNIK: That's right.

MEMBER BIERNER: It's -- the pen and paper form is a document that they make available so we could use it ourselves without paying for

the full calibrated testing with all the background information.

So if you took the ten item instrument and used it in your own clinic, you wouldn't have the advantage of calibrating your scores against all of the independent variables that they have in their database. So that -- so you could still use it, but they do make it available. And that's what's true for all these different body part measures that I was able to find on their website.

CO-CHAIRMAN STILLE: But also --

MEMBER MONROE: Before you stop does that mean that all the demographic data and other data that's in the database isn't available for use in the short form? I think that's what they said.

MEMBER BIERNER: Well, as I understand it, yes. I would probably have to pay to be a -- to subscribe to their service I assume. I mean I'm not speaking to that. I think that's what the website indicts then.

MEMBER MONROE: Okay. So, well, I think you -- thank you.

CO-CHAIRMAN STILLE: David, you had a question?

DR. RESNIK: However, the risk models and all of the variables that are in the risk models, are available on the website. So people who don't subscribe can collect that data. And we provide the coefficients from the risk models so that they can use them.

And then they can compare themselves to basically the average or the predicted values.

MEMBER BIERNER: Yes, that's true. There's a spreadsheet. I pulled those off of the coefficients.

MR. JOHNSTON: And also, the CAT is available to anyone who -- on our website as you know, as requested by the application process. The full CAT survey for each of the body parts is available on the website for anybody to log onto and take the survey and get a risk adjusted measure of the function at that time and the

predicted measure value.

CO-CHAIRMAN STILLE: Okay. David, one last question and then we've got to wrap this up.

MEMBER CELLA: Well, it's not a question I think. But my clarification was just answered. That the bottom line really is whether the short form, and now we just heard the CAT as well, that a provider can use the tool, including the risk adjustment, derive a score and report it without having to be a subscriber. And I hear the answer is yes.

So I think that to me makes it not public domain, but publically available without needing to subscribe. And I think that's what we heard. Is that correct? Is that right?

MR. JOHNSTON: Yes.

CO-CHAIRMAN STILL: Okay. Great, let's go. So, we're going to start to vote on Importance domains for measure 0423.

MS. ALLEN: So we're looking at measure 0423, Functional Status Change for Patients with Hip Impairments. To begin the vote

you will need to point your clicker towards me in my direction.

Please do not start voting until I say start. And I'll go over each slide before you start voting what your options are. The computer will record your last vote. So you can change vote as you desire. But it will only take the last one. Thank you.

CO-CHAIRMAN STILLE: We need a quick clicker tutorial with these new clickers.

MS. ALLEN: So we we're voting on Evidence. You press one or two. I'm going to go through the options.

CO-CHAIRMAN STILLE: Oh, okay.

MS. ALLEN: So, we're voting on Importance, 1A Evidence, rational support of the relationship of the health outcome or PRO to at least one healthcare structure, process, intervention or service. Press one for yes or two for no. Voting starts now.

We have a missing vote. Please -- okay. Results are in.

75 percent yes. 28 percent no. Sorry, 72 percent yes. 28 percent no.

We're voting on Performance GAP. Performance GAP data demonstrate considerable variation or overall less than optimal performance across providers in all population groups, this aspires the use in care. One high, two moderate, three low, four insufficient. Voting starts now.

Zero high, 37 percent moderate, 26 percent low, 37 percent insufficient.

CO-CHAIRMAN STILLE: Okay. So we'll stop now.

MEMBER THOMAS: Insufficient means insufficient information, correct?

MS. ALLEN: Correct.

MEMBER THOMAS: Just can't render a decision?

CO-CHAIRMAN STILLE: Correct.

MEMBER THOMAS: Okay.

CO-CHAIRMAN STILLE: So, are we done?

MS. SAMPSEL: Right. And so the

interpretation here, this is a must pass element. And what this means is we have greater than 60 percent in the low to insufficient category. And you're correct, the insufficient means the Committee doesn't have enough information to further their vote.

I would just ask one last time, you know, the developers do have an opportunity prior to public comment to bring additional information back. And we have the list that Katherine and Sherrie have already provided.

Is there anything else the Committee, you know, kind of direction the Committee would like to give to the developers?

CO-CHAIRMAN STILLE: So do the developers understand sort of the checklist of things that I think came from Committee members?

DR. DEUTSCHER: Can I ask a question? Regarding the insufficient data on GAP. I think it would be good if you could specify, because we did show some information. But apparently it's not sufficient.

If you could specify the specific kind of analysis that you all are actually requesting. So one analysis was mentioned before, showing licensees at the different levels. Are there additional issues? Or types of analysis that you would like us to do?

CO-CHAIRPERSON PARTRIDGE: As one of the reviewers of two of these measures, I would also tell you -- the data that you submitted at perhaps the end of last week didn't reach me in time for me to understand it and digest it. And I suspect that may be true of several of the others.

So I think the staff will work with you on this issue. And we'll be happy to work back and forth with you on the issue through them. It may be some of what you've sent us recently is adequate to answer some of our questions. We'll just see.

MEMBER KAPLAN: Chris, can I just add one quick thing to the developers. It would also help, because we've seen this in other outcome

measures as well. If the components of variation attributable to the patient and then units up to the clinician and then units up to the clinic, can get -- we can get components of variation analysis because that gives us some confidence that.

CO-CHAIRMAN STILLE: Yes.

MEMBER KAPLAN: Yes, it's not all at the patient level. It's not who you see, it's actually what you do. And then maybe at the clinic level who you hire to do that.

So it gives us a little more of the components of variation analysis will help.

CO-CHAIRMAN STILLE: I think it's critically important to translate into policy as well. You know, if I'm a clinician or I'm the boss of a whole lot of clinicians, I want to know okay, where's this variation in care happening and what kind of data do you have to show us where that might be?

DR. DEUTSCHER: So you are referring to hierarchical models that you would like to see

in those -- using those different levels in showing the variance in each level?

MEMBER KAPLAN: There are different ways to do it. But you know, some confidence that some of you know, that the variation of the clinician and the clinic level is actually -- represents a chunk that we would call meaningful.

CO-CHAIRMAN STILLE: And then for the rest of the group, there are a lot of measures in this group. But we want to give all the measures their due if there are differences.

For which measures or for if any, will we have substantially different discussions?

CO-CHAIRPERSON PARTRIDGE: I think what we would like to do is rather then formally vote each one of the other measures, if there's a general sense of the Committee that the issues raised around this measure are going to be raised with respect to the others.

MEMBER SALIBA: It would be helpful just to see the list to answer the question. Thank you.

CO-CHAIRMAN STILLE: Peter had a question.

MEMBER THOMAS: Because these measures are quite similar, it strikes me that I don't -- my guess is that there's not a compelling difference in GAP between some of these other ones. So my question I guess goes to what about the other votes? I mean, I know normally we would stop now and move onto the next.

But because this is so many in one package, would it benefit the developers to go through the process of identifying other strengths or weaknesses in the questions we ask? So that they could prepare all that for the next iteration of this. Or is that just a break in the process? Do you see what I'm getting at?

DR. BURSTIN: I know what you're getting at. I think if there are condition specific issues that are going to come up that might be useful to them as they're preparing the materials back, I think it's useful to them.

MEMBER BIERNER: Yes, I have a

question on the measure 0428, which is what's labeled General Orthopedic Impairment. I wanted -- since we've talked a lot already about specifics that there's different body parts, what is the rationale or who is the audience that you're seeing will use this measure? And when are your clinicians, your therapists, choosing this measure instead of a more specific body part measure?

DR. WERNEKE: This measure would include impairments around cervical, TMJ, thoracic, ribs. Major ones.

MEMBER VAN ZYL: I was actually one of the developer -- reviewers. I'm sorry, not developer. I'm hopped up on Dayquil.

So it seemed to me that the General Orthopedic label was a little bit misleading. Because all of the data that you talked about was really cervical.

And at least when I looked at this, I wondered if this was a measure that would be applicable really to things not mentioned in the

others. Would it make more sense to call this cervical rather than general?

CO-CHAIRMAN STILLE: Can we call up that measure and maybe we can look at that? 0428. Sure, why don't you answer while we're looking for it. That's great, thanks.

MR. JOHNSTON: Yes. The reason we presented the cervical data was because it was the predominant data in that -- or the predominant impairment group in that database.

MEMBER VAN ZYL: Right.

MR. JOHNSTON: I think it was 70 percent of the data that was the cervical. But we still had the other 30 or 40 percent that were other body parts. So we elected to keep it.

You know, I think that the progression of this would be to separate out a cervical measure from this. Because we do have those other generalized orthopaedic impairments that need to be measured by something. And we have people participating in it for that.

But because of the large number of

cervical patients, I think we need to have a cervical measure. And we're actually working on approving one for our future submission.

MEMBER VAN ZYL: So you're already thinking about separating the cervical out explicitly. Okay.

CO-CHAIRMAN STILLE: So then to go back to Peter's question. What else can we do to be helpful in terms of either voting, giving the developer some more data that they can take back?

MEMBER NEUWIRTH: I guess this is a question maybe more for us. But maybe for the developers as well. I guess I'm thinking about you know, sort of feasibility and usability, especially coming from an integrated system. Where the body part specific surveys to some extent make sense, but then also, when I look across these different instruments, they look very similar at least in terms of the tenth item one.

And so I'm thinking about -- and maybe, I haven't looked comprehensively across

all of them. But I guess I'm just wondering you know, when we're thinking about cost and we're thinking about also you know, ease of application and desire for spread, are all these different instruments really going to make sense in practice?

And also the time that it's going to take you know, regularly to review these and so on. And then the cost associated. And I also feel like there's a question in my mind about you know, as patients, I think we hear over and over again that they're not a body part, that they're a whole person.

And so all of that to me questions sort of the distinctions between these different instruments and how useful and valuable it is to spend this much time and all these you know, developing these individual ones. When if we had a sort of holistic approach that might actually even better serve our patients.

So and that might be you know, moving forward, I think you mentioned Helen, that

there's a desire to collapse some of these. So I would look to the developers to maybe speak about that as well. But also for us as a Committee.

DR. DEUTSCHER: I think one of the answers are given by the different analysis we've done. And I'd like to give a simple example just to illustrate that.

Some of the measures that are very similar, actually they use the same items are hip, knee, foot and ankle. Coming from the lower extremity function skill.

But when we analyze the data for differential item functioning, which means that the patients might perceive different items having different difficulty levels. If we do not take and we found differential item functioning for example, for these measures.

So what that means is that a specific function might be perceived having -- or representing a different difficulty level whether I have a hip problem or a knee problem. And when those differences are found and they're

significant, so the measure works better if we recalibrate the difficulty level of the items for each of these body parts.

So it just makes these measures being more precise, more responsive. And that's the reason why they were separated.

DR. WERNEKE: And as a clinician, although they're coming up let's say with the back problem, you really have to address the patient's functioning in their ADLs and at work, et cetera. So if you focus on back, then you have to incorporate it into the total body for the purpose of improving their function and their perception of their function.

So I see us treating the whole body. We're just not treating the knee. But we want to know how that integrates or plays with their role in ADLs, work, et cetera. And that's important.

And if you do not do that and all you do is focus on their low back, you will not improve their quality of life or improvement in their self-report outcome. You won't see that.

So unless you address their -- the total package during that episode of care, you're not going to get higher patient self-report outcomes. So you can't just focus on one impairment during the treatment episode. Yes, but the survey captures that.

DR. RESNIK: Well I think one of the important points that I'd like to just reiterate is if there is differential items functioning and difficult -- differential difficulty for people with different impairments answering the same questions that if we don't separate the measures, the scores will not be accurate.

And then we'll have people basically answering on different metrics. And then we won't be able to compare them. So as much as it would be nice to have the universal measure to compare across all impairment types, it really wouldn't be valid.

CO-CHAIRPERSON PARTRIDGE: Peter, could we go back to your question. Is -- are you suggesting that we vote 1A and 1B individually

for each measure as a process?

MEMBER THOMAS: Sorry, thank you. I'm suggesting I guess for this measure that we go down the line of validity, reliability, use, and feasibility.

CO-CHAIRPERSON PARTRIDGE: Through 2A and B and 3 and 4?

MEMBER THOMAS: Just to give them a sense of whether there might be --

CO-CHAIRPERSON PARTRIDGE: Okay. Thank you very much.

MEMBER THOMAS: One real other weakness they could work on in the meantime so that they have to keep --

CO-CHAIRMAN STILLE: Yes. That's what we're talking about back here.

CO-CHAIRPERSON PARTRIDGE: All right. Okay. So the question before us is assuming that the measure had passed 1B, let us go on and vote for this measure, but thinking it's the others?

CO-CHAIRMAN STILLE: Yes. We're thinking probably between the group.

CO-CHAIRPERSON PARTRIDGE: To 1C? Maybe?

CO-CHAIRMAN STILLE: Liz, did you have something to say before we -- you had your thing up.

MEMBER MORT: I did. And if I were the developer, I would be wondering after having been approved by NQF a couple of times, and then getting this response, did something change? Or did our criteria change?

CO-CHAIRPERSON PARTRIDGE: The answer is, yes. The developer has indicated, this is first of all modified from the approver's -- I'm right, yes?

MS. SAMPSEL: Correct. And I think if Ben -- well what Ben had summarized is this was originally a process measure. They've moved it to an outcome measure. So that was a significant change. And also the age range from 18 to 14.

MEMBER MORT: Well the PM part I get. I mean the PRO-PM. But some of these basic things like evidence would have been before the

NQF group approving it in the past. Anyway, it just seems, I might be confused if I were them. Okay.

MS. ALLEN: Voting is open now for 1C, High Priority. One high, two moderate, three low, four insufficient information. Voting starts now.

We're still waiting on a vote. 37 percent high, 58 percent moderate, five percent low, zero percent insufficient information.

Voting on Reliability. One high, two moderate, three low, four insufficient. Voting starts now.

11 percent high, 21 percent moderate, 37 percent low, 32 percent insufficient.

Voting on Validity. One high, two moderate, three low, four insufficient. Voting starts now.

11 percent high, 11 percent moderate, 47 percent low, 32 percent insufficient.

Now we're voting on Feasibility. One high, two moderate, three low, four insufficient.

Voting starts now.

37 percent high, 42 percent moderate, 16 percent low, five percent insufficient.

Voting on Usability. One high, two moderate, three low, four insufficient information. Voting starts now.

We're still waiting on a vote. 21 percent high, 26 percent moderate, 37 percent low, 16 percent insufficient information.

CO-CHAIRMAN STILLE: So, just observing as the data have gone in. I think if a lot of the insufficients could be converted to highs or moderates, you know, the numbers would be there.

So I think this is much more a plea for more information than anything else.

DR. RESNIK: We haven't addressed usability. As far as a recommendation that I am aware of.

CO-CHAIRPERSON PARTRIDGE: Linda, we did. When we were voting on 3 and 4 we did.

CO-CHAIRMAN STILLE: Yes.

CO-CHAIRPERSON PARTRIDGE: And they passed.

CO-CHAIRMAN STILLE: Yes.

DR. RESNIK: I see.

CO-CHAIRPERSON PARTRIDGE: Okay. Thank you all. And let's be back in ten and then we'll take up 26 -- no, yes, it's 02624.

(Whereupon, the above-entitled matter went off the record at 11:01 a.m. and resumed at 11:23 a.m.)

CO-CHAIR STILLE: Welcome back. One of my career mentors ten years ago acquainted me with the phenomenon of the miracle of the agenda, which basically says that no matter how crazy agendas get during the meeting, by the end of the meeting everything else ends up being discussed on time, most of the time. So, we have accomplished a miracle.

Just real quickly just in case there was any confusion, the FOTO measures have been discussed. We feel like everything is adequate to inform everything we need to move forward on

that.

Do you want to do logistics and housekeeping then, and then dive into the next measure? Okay.

MS. THEBERGE: Sure. So, we have had a bunch of questions about whether there is going to be future meetings. At this time, we don't have a Phase 3 for this project funded, but it's possible it's going to happen.

So, we do not at this time have another in-person meeting scheduled. We will keep you posted if and when that changes.

We do have a call scheduled for next week, and then we will have a call after comments. And we will probably be having an additional call at some point to deal with some related and competing issues, but, you know, we'll keep you all posted by email.

There will be some surveying on availability and all that, but -- and there may be no need for the call next week. We'll have to kind of see how today goes and how much we get

through, but we'll keep you all well-informed on scheduling.

CO-CHAIR STILLE: Okay. So, we're going to proceed then to discussion of Measure 2624, the functional outcome assessment from CMS. The measure developers are here and the discussant, Katherine, will probably take the lead on the discussion, I assume.

Okay. Go ahead.

MS. SAMPSEL: And before I turn it over to Sven, I think there are a number of folks on the phone from CMS, correct, and perhaps Quality Insights as well.

So, if you could just announce yourselves real quick so we know who's on the phone?

MS. AUTREY: Good morning. This is Sophia Autrey calling from CMS.

MS. SOMPLASKY: Good morning. Anita Somplasky from Quality Insights.

MS. LUCAS: Jane Lucas and Jeannette Shrift from Quality Insights.

MS. GOERTZ: Christine Goertz with Quality Insights.

MR. REZEK: This is Gary Rezek with Quality Insights.

MR. BERG: Good morning, everyone. And although Sarah after this morning's earlier session gave me the opportunity to just run away --

(Laughter.)

MR. BERG: -- we're really, really happy to be here. And on behalf of the Centers for Medicare and Medicaid Services and the measure's developer, the Quality Insights of Pennsylvania, it's my pleasure to introduce to you NQF 2624 Functional Outcome Assessment for consideration of NQF endorsement.

This measure was actually initially developed in 2008 and was implemented as part of the Physician Quality Reporting System in 2009. An effort to fill a gap in reported measures that addressed clinical strategies that were relevant to the chiropractic community.

Since its initial implementation, the measure's use has been expanded to include physical therapists and occupational therapists as well.

NQF 2624 measures the use of a standardized functional outcome assessment tool by eligible providers to identify deficiencies and provision of a care plan that addresses the deficiencies identified.

Performance is assessed for all visits for patients aged 18 years and older, and reporting is required for each visit for patients seen during the 12-month reporting period by way of administration -- administrative claims or a registry.

As you all know, standardized outcome assessments, questionnaires or tools are a vital part of evidence-based practice, and outcomes measures along with other standardized tests and measures used throughout an episode of care are being -- as part of a periodic reexamination provide information about whether predicted

outcomes are being realized.

Despite the recognition of the importance of outcome assessments, questionnaires and tools, evidence still suggests that their use in clinical practices is still limited.

In addition, frameworks, guidelines by associated specialty societies support the documentation of the use of assessment tools, as well as documentation of a plan of care on each visit.

A need for improvement in care provided using this measure is evidenced by an average provider performance rate of 80.9 percent in 2012.

Differences in performance rates based on various demographic traits, for example, statistically significant performance gaps between urban/rural, male/female, non-white/white, ethnicity and age groups.

And so, we believe this measure addresses the importance of utilizing validated functional assessment tools to monitor the

patient's status and initiating adjusting care plans as appropriate.

So, we thank you for the opportunity to present today. Thank you for your consideration of endorsement, and we look forward to the Committee's questions.

CO-CHAIR STILLE: Great. Thank you.

Katherine, all yours.

MEMBER BEVANS: Yes, I just have a few comments and questions before we ask other people to add.

As this is one of, I think, the first, if not one of the first process measures that the Committee has evaluated, I'm wondering if you could provide a rationale for why and what evidence is there around the use of standardized functional assessments and care planning and outcomes, and what the evidence around that actual documentation, how that changes, whether or not it is associated with better outcomes for patients.

MR. BERG: Sure. And we -- I think we

have a member of our technical expert panel online who probably would be best able to answer that question, Dr. Goertz. So, if she could answer that question?

MS. GOERTZ: Yes. Thank you.

Could you please repeat the question?

MEMBER BEVANS: Yes, the question is --

MS. GOERTZ: I just want to make sure I understand.

MEMBER BEVANS: Yes. Sure. I'm wondering what prior evidence not necessarily for application of this measure, but prior evidence and research suggests that documentation of a standardized -- use of a standardized functional assessment and care planning, what that means for patient outcomes.

Is there an established link between the activity that the process measure is assessing and improved patient outcomes?

MS. GOERTZ: Right. There is definitely an established link between the care itself and the outcome measure.

The type of standardized tool is commonly used to assess the outcome of chiropractic care both in clinical practice and in research situations.

It's less clear to what extent actually the measure --- what component the measurement itself is contributing versus what component the care is contributing.

MEMBER BEVANS: Okay. So, if I'm understanding that correctly, there's not necessarily very strong documentation of the linkage between the actual recording of the use of this tool and patient outcomes; is that correct, or am I missing something?

MS. GOERTZ: In the chiropractic population, not that I'm aware of.

MEMBER BEVANS: Okay.

MR. BERG: So, if I understand your question, it's the link between reporting of the tool, not the link between the use of the tool.

MEMBER BEVANS: If that is what the process measure is, in fact, getting at, right?

MR. BERG: And I don't believe that we've assessed the effect that reporting has on performance.

MEMBER BEVANS: Okay. And I assume, you know, really a question to our leaders then, this is an issue in evaluating a process measure around the importance of that measure.

I notice also that with regard to reliability, the inter-rater reliability was fair only and I am wondering if there has been any attempt to kind of mitigate how the data are actually collected to improve inter-rater reliability.

MR. BERG: Sure. And I'll have Hiral talk to that issue.

MS. DUDHWALA: Yes, that's something that we observed as well. There was fair reliability when we compared our independent reviewer and our claims, what had been reported.

And what we found after we looked at that information, was what was lacking was a clear documentation of the outcome assessment

tool in the claims.

So, you know, as a response to that, we did take this back to our technical expert panel team and we did identify this and clarified the specification so that this is very clear, you know, for the providers that, you know, the name of the tool does need to be documented, because that was --- that was what --- really what was a big gap in that area. So, we did update that specification to note that.

MEMBER BEVANS: Okay. And the documentation, I may have missed it, but did that activity actually significantly improve the inter-rater reliability?

MS. DUDHWALA: So, that update just happened in the 2014 specification. So, further testing to see how that improved would happen this year.

MEMBER BEVANS: Okay. So, we're not quite sure there about that.

MS. DUDHWALA: Yes, we're not quite sure, but we did notice that was the issue that

had been showing up.

MS. SAMPSEL: So, Katherine, can we do importance first this time and --

MEMBER BEVANS: Yes.

MS. SAMPSEL: -- kind of focus so we can --

MEMBER BEVANS: Absolutely.

MS. SAMPSEL: -- so we can focus the conversation and vote? Thank you.

CO-CHAIR STILLE: I had a question about performance gap, which is actually pretty much right what's on your screen right now is that the median for providers although there is not very many providers that were reporting, was a hundred percent. And how do we kind of figure that out to maximize the value of this measure?

MR. BERG: Is Gary online? Does he want to answer that question? Gary is our statistician.

MR. REZEK: Well, yes. I would -- I can address that mainly by a point I try to emphasize. We're looking at the providers who

chose to report this on claims. So, it is -- it's a small proportion of the total eligible population of providers who could have reported the measure. So, do have to take performance data with a grain of salt.

I think, you know, the median could be, you know, since it's a sort of self-selected group of providers who are reporting this, that the performance is -- maybe it's the high performers who are reporting, but we don't know that for sure.

So, although the median rate is a hundred percent, we do see, you know, I believe -- and I'm not looking at the webinar. I'm sorry. I'm remote here, but I believe our average performance was something in the range of 80 percent. And we do see a lot of variation sort of in the bottom 50 percent of reporting providers.

CO-CHAIR STILLE: Right. And I think only like four percent of providers reported, too. So, yeah, I think your idea of some

selection bias in that initial sample is right.

MR. BERG: We do have an update in terms of -- and it's not part of the package here, but in terms of the number of providers or percentage of providers who are now using the tool as well. In preparation for this meeting, we went back and looked for the most recent data.

So, in 2013 there has been an appreciable increase in the number. And of providers that made application to attest to meaningful use, the utilization now is about 25, 26 percent.

CO-CHAIR PARTRIDGE: And following up on Chris' question, in your 2013 data did your median and your percentiles change?

I mean, that 50th percentile, a hundred percent is kind of a big flag.

MR. BERG: Right. And, again, we just started to pull -- we just started to pull that data. And that's -- so, we haven't looked specifically at that in the 2013 data, because the data is just preliminary and not mature yet.

CO-CHAIR STILLE: Yes, David.

MEMBER CELLA: I'm sorry if I missed this, but to Katherine's point we're talking about importance, right?

So, there's an "and" in this numerator which is not just the documented functional outcome assessment, but a care plan that's based on the identified functional outcome deficiencies.

Is there a way, I mean, is that a true and, meaning -- because it seems to me that if there's a care plan that's tied to the functional assessment, in my mind that would become more important than if they just did the assessment.

So, in the way that this is collected and reported, is the link between the assessment being done and the care plan being provided, clear?

MR. BERG: That's the intent of the measure is for that to happen. And, again, the potential weakness, I think, has already been shown in terms of the difference between the

collected data, you know, the reliability data itself.

And so, the one thing that I can say is that when we compare the abstractors information to our own abstractors to determine an inter-rater reliability between the abstractors and then ourselves, that reliability went up into the 80 percent range itself. And so, an understanding amongst the abstractors and ourselves as to what was required for this measure was found.

However, we recognize the need to go back and to reevaluate following the changes that we have made to see if the increase in reliability has been accomplished as well.

CO-CHAIR STILLE: Sherrie.

MEMBER KAPLAN: I probably am going to win the prize for the most confused in the room, because I'm kind of confused about what -- when I first read this measure, I didn't know what we were actually being asked to endorse, because it says that a suite of, quote, standardized

functional status measures can be administered.

And then it says, unless I missed something, Kevin, because you can help me --- I read this over four times. I really did try to get this right.

And then -- and so, if you do it, it's zero/one at the patient level. So, it either was done or it was not done, but there's a whole host of different things that could be administered.

And then you have to interpret it --- you have to score it, interpret it correctly and formulate an appropriate functional impairment reduction plan. And that's what counts as you get a one. Then if that was done, you get scored.

Just to stop there, is that what you ---

MEMBER BEVANS: That was my understanding. Please, correct us. It's all or nothing kind of scoring system, right? Is that correct?

MR. BERG: Yes, that's correct.

MS. DUDHWALA: Yes, you have to -- you have to pass both parts to meet the measure, to pass the measure.

MEMBER KAPLAN: So, then my question is, who said somebody administered the right one? Was it sensitive and specific to the problem under consideration for improvement, A? B, how are you evaluating the functional improvement plan? Has that got some levels or tiers of it, or it was just done or not?

Is this a documentation measure, or is this a quality of care measure?

MR. BERG: Yes, it's a documentation measure.

MEMBER KAPLAN: Okay. So, now, NQF, you have to help me understand does that fall under the rubric here of a --- it's a documentation measure. So, you're documenting --

CO-CHAIR STILLE: So, it's a process measure, you know.

MEMBER KAPLAN: Okay. You're just documenting that it was done. But then if it was

done right, what happened?

I mean, how do you know if it was done correctly? The appropriate measure was applied, the appropriate care plan was applied. It's in the record and you can conceive of some EMR results that are just going to not let you go any further until you say yes or no and get a score right away as opposed to people who are EMR, some EMR and don't.

MR. BERG: This is a process measure. And so, that's not the purpose of the measure to evaluate whether it was done correctly or not.

And we recognize that process measures are beginning to fall out of favor at this point, and we really are looking for more on the outcomes-based type of measures.

This measure was developed, again, back in 2008 and implemented in 2009 by CMS as part of the PQRS program.

And so, this actually was supposed to come to this committee three years ago. And so, that's been -- that's been a bit of a delay. And

so, we understand the limitations of process measures, but it wasn't really built at the time it was built to evaluate the correctness of what was done.

MEMBER KAPLAN: Can I ask one follow-up question before we leave that?

MR. BERG: Sure.

MEMBER KAPLAN: So, if it's zero, you don't know if it wasn't done, or if it's just missing. That's what the nature of documentation is, right?

MR. BERG: That's correct.

MEMBER KAPLAN: Okay.

MR. BERG: That's correct. Yes.

MEMBER KAPLAN: So, then if you have a bunch of these things being done and now the median score --- this is a follow-up on Lee's point.

If the median score is a hundred percent, at what point do you retire this, because it's no longer varying or getting ceiling effect problems.

MR. BERG: We probably don't have enough data to know that because utilization was so small at the time.

CO-CHAIR STILLE: Right. Right. Yes. Once you have more complete data, you'll know is it still useful or not.

MR. BERG: That's correct.

CO-CHAIR STILLE: Sherrie, I might add that you know what you don't know, the rest of us don't know what we don't know. So, that's why you're making comments. So, thank you.

MR. BERG: And one thing you might argue is because the utilization up to this point or at least until recently has been so low, that in and of itself is perhaps evidence of a gap itself that needs to be filled.

CO-CHAIR STILLE: Peter.

MEMBER THOMAS: I'm having trouble understanding the 25 percent figure that you quoted of 2013 data.

Are you saying that providers in --- 25 percent of providers did this?

MR. BERG: So, there are actually two numbers. There are actually the number of eligible providers which we calculated at that time to be somewhere over --- it was over 90,000 providers.

However, of those providers, not all of them at that time had signed up to participate in the meaningful use program.

And so, of a smaller percentage of those that had signed up for meaningful use, of that population we found that 25 percent were using this measure.

MEMBER THOMAS: Okay. Using the measure, but not necessarily --- it doesn't necessarily correlate with whether or not they're doing this.

MR. BERG: That's correct.

MEMBER THOMAS: I mean, frankly I find it astounding that you would go to a PT or an OT or a chiropractor and they wouldn't assess your functional level and develop a plan of care.

I mean, what else would you be doing

if you weren't doing that to serve a patient's needs?

MR. BERG: This measure is designed for that to occur at each visit as well. So, that was part of the measure so that there would be an ongoing assessment of the functional status of the patient.

MEMBER THOMAS: Okay.

MR. BERG: And adjustments as necessary to the care plan.

MEMBER THOMAS: Thanks.

CO-CHAIR STILLE: Ann.

MEMBER MONROE: I'm struck by the disparity discussion that's here. In terms of the difference, I assume now it's in documentation or completion of this assessment, right?

Is that what the disparities refer to? I mean, they're very high, I think. You talk about statistically significant for gender and age and even larger differences between urban, rural providers and patient race, ethnic group.

So, how do you look at those disparities? Are they of sufficient significance to you that you think this measure just isn't being adapted, or what do you think that's saying?

MR. BERG: Dr. Goertz, do you have an opinion on that? Obviously, you know, to me when I look at the amount of disparity that's there, it certainly means that there is a need for the information to be there. There is a need for the information to be considered.

The thing that I thought was interesting, though, is the disparity went in the opposite direction in terms of the race population --

MEMBER MONROE: Right.

MR. BERG: -- as I thought it would. And I was very surprised by that. And I don't really have an explanation for that, because it appears that the performance was better in the minority groups than in the white Caucasian group.

MS. GOERTZ: Yes, and I'm not actually able to answer that question. The population that goes to a doctor of chiropractic is not as diverse as populations that may go to some other providers.

They tend to --- chiropractic patients tend to be Caucasian. They tend to be in a little bit higher socioeconomic status. And that tends to be the people that we attract in our randomized clinical trials as well.

To date, there hasn't been a study that has been sufficiently powered to look at differences in outcomes based on some of those criteria, though.

We're currently conducting a large-scale trial in the Department of Defense that should give us that data for the first time.

CO-CHAIR PARTRIDGE: Ann, I think also it's possible --- Mr. Berg, you said this is a meaningful use measure.

Part of the question could obviously be whether or not you have access to funding for

an EMR, a medical record system could turn on where your practice is located whether you're affiliated with the hospital and so on. So, that may account for some of your disparities, too.

MEMBER MONROE: I'm struggling with how to assess that. I mean, it feels very significant to me and what does that say about the measure? How do I interpret that thought in the face of what our task is?

I mean, does that make the measure less effective, score it more, I mean ---

CO-CHAIR STILLE: I think you need to take it in the context of these are the data that are available at this point.

MEMBER THOMAS: Doesn't it suggest that there is a need for the measure so that those that are not doing this and they are patients that are not experiencing this process are, therefore, higher likelihood that they would be exposed to that as a result of having this measured and tracked and isn't that what that means?

CO-CHAIR STILLE: On the end.

MEMBER LINDBERG: Thank you. I'm actually a fan of process measures. I think that, you know, the most maybe overused example of the process measure is the preoperative antibiotic, you know.

And the fact that once we started measuring that, doctors, hospitals, everybody, they're checking the box. Yes, somebody got the antibiotic.

I think this, unless I'm off here, it seems to me like this makes good sense to make sure that people check this box and that they've done this for each of their patients and they do it regularly as opposed to maybe once.

CO-CHAIR STILLE: Sort of a mantra in the quality field is not documented, not done. So, okay.

One more comment, and then I think we need to vote on importance. Sherrie.

MEMBER KAPLAN: I just had a quick question for clarification. Do the measures in

the suite of standardized measures have to be NQF-approved, or is -- they don't.

So, this is an NQF-approved measure of things that NQF has --

MR. BERG: I would say these are not necessarily measures that we're using. They're just the use of a functional tool, functional assessment tool.

So, yes, it would be correct they're not necessarily NQF-approved functional assessment tools.

MS. SAMPSEL: Well, and we want to clarify NQF doesn't review or endorse tools. Those don't come under the purview, just the measures that might be the result or the outcome of the tool.

MEMBER KAPLAN: I understood that part. I was just confused about the link. And then, still, the suite of acceptable measures is listed. Somewhere there's a long list for the coders to say it is or it isn't.

Nobody can just make up their own

little whatever or use something in children that's been used in adults and by the way, is pediatrics part of this, or is this an adult measure?

MR. BERG: No, this is AJT and older.

MEMBER THOMAS: Is there any setting other than outpatient that this would apply or that this is used or could be used?

MS. DUDHWALA: It's just currently outpatient setting at this point.

MEMBER THOMAS: Okay.

CO-CHAIR STILLE: Okay. Can we vote on importance?

MS. SOMPLASKY: This is Anita Somplasky. Can I just clarify something? This is not a meaningful use measure. It's a PQRS measure, which has made it a little bit more difficult to ascertain, you know, the actual documentation because you have to wait to see who has reported through the PQRS program and then be able to ask for a sampling of those to see if the documentation is present.

CO-CHAIR PARTRIDGE: And if this is not a meaningful use measure, then I take back what I said about whether or not you had access to an EMR.

MS. ALLEN: So, now we are voting on evidence. One, high. Two, moderate. Three, low. Four, insufficient evidence. Five, insufficient evidence with exception.

Voting starts now.

CO-CHAIR STILLE: And what does exception mean?

MS. SAMPSEL: Exception would mean that --- all right. So, and we had a big talk about this yesterday on if we should even have this category anymore, but basically what it means is not enough information was provided, but this is an important enough concept and you think the evidence is there based on feedback from other members of the Committee that you would give it an exception.

So, that Five category would actually, you know, so, let's say that, you know, you had

54 percent in one and two, and seven percent in Number 5. That takes it over the 60 percent. So, it would be above the gray zone and it would pass, if that makes sense. It counts as a pass.

MS. ALLEN: Voting starts now.

(Voting.)

MS. ALLEN: 11 percent high. 47 percent moderate. 16 percent low. 16 insufficient evidence. 11 percent insufficient evidence with exception.

MS. SAMPSEL: So, we do move -- we continue to move forward with this one.

MS. ALLEN: Voting on performance gap. One, high. Two, moderate. Three, low. Four, insufficient information.

Voting starts now.

(Voting.)

MS. ALLEN: 21 percent high. 63 percent moderate. 16 percent low. Zero percent insufficient.

Voting on high priority. One, high. Two, moderate. Three, low. Four, insufficient.

Voting starts now.

(Voting.)

MS. ALLEN: 21 percent high. 63 percent moderate. 16 percent low. Zero percent insufficient.

CO-CHAIR STILLE: Okay. Psychometricians, have at it.

MEMBER BEVANS: A quick question about reliability or the specification, I guess. And this kind of gets back to Dave's comment as well.

I think that in terms of operationalizing this measure, choosing from a list of potential outcome measures or standardized tools makes sense as part of the first element of the process measure, but could you help us to understand a little bit more about how the second element is operationalized?

How do we know that a documented care plan is based on the identified functional outcome deficiencies?

It is one thing to be able to document a care plan was generated, but the qualitative

element of that definition is based on the functional deficiency. It's trickier.

And so, I'm hoping you could help us to understand how that's operationalized for the purposes of coding.

MR. BERG: Dr. Goertz, do you want to answer that question?

MS. GOERTZ: I can talk a little bit about how it's operationalized or what our thought was when we were putting that together. I would not be able to answer how it's actually operationalized for coding purposes.

I can talk about our intent and the training, but I --- so, we added the care component, I think, in the second or third year after the measure was developed because we wanted to make sure that there was a direct link between quality of care and -- we have trained the doctors of chiropractic most familiar with the training that's gone to them about the importance of the measure itself and that it be linked to a care plan.

And I believe that they are instructed to record the date in which the care plan is developed and the dates on which it is modified, but I wasn't involved in any sort of an audit that showed the extent to which that's actually occurring.

CO-CHAIR STILLE: I'm a little bit concerned just as this continues to roll out, about the reliability in as reported in a much bigger sample.

Sometimes it's hard to tell if there's a care plan in a medical record, for example. And so, I think it's going to be really important to get follow-up data about how accurately can we tell whether this stuff is there or not.

It's hard to tell with the sample that we have right now, but it's hard enough to get anything out of an EMR. And sometimes care plans can be a little bit of a weird part of that.

MEMBER KAPLAN: Can I -- the reliability issue is just reproducible. And that's the agreement, you know, somebody looking

in this case at the same documentation or lack thereof and reproduce it, but just being -- what I understand is just if somebody said it was there, a care plan was there at all, it doesn't matter. Those two things could have been completely independent of each other.

Somebody did the functional status assessment. Somebody else wrote a care plan. Bingo. Both of them are there and the measure is satisfied. That's correct, right?

It doesn't matter if they were linked. They were just there. Somebody went through the record. Bingo, I found one. Bingo, I found two. Bob's your uncle.

MR. BERG: That's how the measure is designed.

MEMBER KAPLAN: So, then the question for reproducibility is did somebody else looking at that same information get the same answer? And what I understood from you is, not so much.

If you look at the medical record, if you go back and abstract the medical record and

try and compare it with the claims data results, you don't necessarily get too much agreement between those two sources and the same information; is that right, or no?

MR. BERG: The initial data would suggest that, again, we want to go back and re-look at that following the clarification that we gave after that initial data was obtained.

MEMBER KAPLAN: So, then the issue of who's right, you know, which is a validity issue, because then is it accurate, you know, that becomes then the validity question that we don't know the answer to yet.

MEMBER BIERNER: Well, it would be rather easy to have collected data that would say, this is the measure I used.

I understand you have a suite of previously validated measures like Oswestry or whatever, but it could have been a checkoff box. I used for this patient Oswestry or I used neck disability index, whichever, and identified functional outcomes or functional goals for the

treatment and that could have been specified.

I mean, it would be pretty easy to specify that in a general way. You could say we're, you know, these are the deficiencies in function.

So, I just have a lot of problems with this that we're -- that this measure which hasn't yet been widely used among the community of providers of chiropractors, it could be so much better and collect actually more useful information.

Otherwise, you're just saying there's a piece of paper in the record and nothing about whether it really relates to the functional deficits outlined in the tool they use, which just seems like a waste of time.

MEMBER BEVANS: I think this point about the connection between use of the standardized tool and the actual care plan whether the care plan is informed by results of the tool is really a key issue.

And I think that for me, I could

really better understand and make a judgment about this if we could know specifically how the observation was operationalized, you know, what exactly are the coders looking for. So, that information could be key.

If Sherrie is correct in saying what we're looking -- what the coders are looking for are did you use a measure and do you have a care plan, then if that's correct, then I think that the description of this measure in the document is a bit misleading because what it says is a documented care plan based on the identified functional outcome deficiencies.

That may or may not be true, you know, if all you're looking for is use of a tool and have a care plan, but it's hard to say maybe that's not actual, you know, how the measure is operationalized. So, more information would be helpful.

MR. BERG: And there are two sources of data for the calculation of the measure. One being claims data where there really is no way to

code that connection between the two, and then in a registry.

And the number of -- the percentage that is in registry as opposed to claim data is much smaller than that. But as a claims-based measure, I don't see a way to make that linkage based on the way claims are coded at this point.

MEMBER BEVANS: So, it would be important to respecify the definition?

CO-CHAIR PARTRIDGE: It's been a long time since I had to read claims. But as I remember, G codes are not payment-based. They're kind of additional information stuff.

MR. BERG: That's correct.

CO-CHAIR PARTRIDGE: And they won't be in any way detailed. What we determined was this person needs work on exercising his shoulder, right?

CO-CHAIR STILLE: Other questions?

Dave.

MEMBER CELLA: So, I just wanted to follow up on that point that Katherine was

bringing back which is exactly why I asked the question when we were voting on importance, was it based on part of the numerator. And you said that's the intent.

And so, I voted in favor of it being important because of that word "based on." I mean, that literally, to me, made the difference.

And then Sherrie, I think, illustrated that, and I think you've confirmed, that basically the way it would be done today, there would be no way to confirm based on.

So, I think at least from my perspective, what we have is a case where this is an effort to move a process measure into more of an outcome-like measure and more of a care --- a real care-based measure a little bit analogous to antibiotics.

I mean, people will do the right thing and the right thing should be tied to that tool and not just I've got to write a care plan because, you know, I have to.

So, it's important to make that based

on link, but I think what we're hearing is we're not sure how that's going to happen unless the system changes.

So, the reason I'm maybe belaboring this point is that I don't know how that means we should vote going forward. Because on the one hand, I personally as a member believe that it's important to have that based on link and that this should be encouraged and promoted in some way, but I don't know how when it leaves this meeting and then goes into use would we be favorably, you know, stamping something that will continue to be rolled out as Sherrie illustrated, you know, you do A, you do B, you get the one, you're in the numerator.

That would trouble me unless there was some way to get some teeth into that based on link that really to me is the core.

CO-CHAIR STILLE: Liz.

MEMBER MORT: I think another way of saying that is that this is very game-able. And game-able is a risky methodology, because that's

when you run into the unintended consequences associated with putting in non-linked plans or other such things.

And I know that some organizations are really taking a pretty active stance against promoting measures that are game-able.

I mean, perioperative antibiotics there's fraud. You can lie and you can check the box that said you did it.

But if you do it, there is a very -- and it gets into the patient, that's not game-able. You've given the antibiotics as you should, but you could put a care plan down that clearly would be meeting the metric, but not necessarily delivering the care that would be right for the patient.

CO-CHAIR STILLE: So, it might be helpful for the developers to talk about kind of how this is measured, because there's some questions about that.

MR. BERG: Gary, do you think you can answer that question? Are you still there, Gary?

MR. REZEK: Ye, and I'm not exactly sure how our testing addressed that issue, the issue how this is essentially being implemented at the provider level and how they would define a care plan and if that explicit connection has been made between the outcome deficiencies.

We don't really have that answer in our data.

CO-CHAIR STILLE: Okay. Great. Let's vote.

MS. GOERTZ: Could I just say one thing? This is Christine Goertz.

CO-CHAIR STILLE: Yes.

MS. GOERTZ: Before you vote.

CO-CHAIR STILLE: Yes, please.

MS. GOERTZ: Oh, thank you.

I'm both a clinician and a scientist. And as a scientist who does randomized clinical trials for a living I completely understand this discussion and the need for data and that clearly links the care plan with the collection of the outcomes data.

As a clinician, I can tell you that when you have that data, you link it to the care plan. It's not something that you would necessarily gain as --- if you have that data collected, you would just naturally link that to your care. It would be data that you wouldn't just ignore. So, I would just ask that you keep that in mind as you're voting.

While I understand the need for data and I'm thinking of ways right now that we might be able to educate our providers to make sure that we are, in fact, able to make that link and that we do it in a way that's auditable, I would -- I just would like you to think about it just a little bit from a clinical perspective and all where that link would just naturally be made.

MEMBER KAPLAN: Can I respond to that, because the one thing patients complain about almost uniformly is they complete these forms in the office at intake of review of systems and they carefully complete them and often are frustrated because they may not have them

finished before they go to see the doctor. Do you have any diabetes? Blah, blah, blah. And the doctor systematically ignores all those data.

So, I don't know that just because somebody filled out the form, that it actually got integrated into -- effectively into care plans without some extra steps.

CO-CHAIR STILLE: Yes. A quality metric for quality of a care plan is so badly needed, I think.

Anyway, let's vote.

MS. ALLEN: Voting on reliability, which includes precise specifications in testing. One, high. Two, moderate. Three, low. Four, insufficient.

Voting starts now.

(Voting.)

MS. ALLEN: Zero percent high. 53 percent moderate. 21 percent low. 26 percent insufficient.

MS. SAMPSEL: So, this is considered in the gray zone, but we still move forward to the

next vote.

MS. ALLEN: Voting on validity including specification consistent with evidence, testing, exclusion, meaningful differences. One, high. Two, moderate. Three, low. Four, insufficient.

Voting starts now.

(Voting.)

MS. ALLEN: Zero percent high. 42 percent moderate. 32 percent low. 26 percent insufficient.

CO-CHAIR STILLE: Brief comments on feasibility.

(No comments.)

CO-CHAIR STILLE: Anyone? Should we vote?

(No comments.)

CO-CHAIR STILLE: Let's vote.

MS. ALLEN: Voting on feasibility. One, high. Two, moderate. Three, low. Four, insufficient.

Voting starts now.

(Voting.)

MS. ALLEN: 16 percent high. 58 percent moderate. 26 percent low. Zero percent insufficient.

CO-CHAIR STILLE: Comments on usability. There were a few from before.

(No comments.)

CO-CHAIR STILLE: Anything? Okay. Should we vote? Let's vote.

MS. ALLEN: Voting on usability. One, high. Two, moderate. Three, low. Four, insufficient information.

Voting starts now.

(Voting.)

MS. ALLEN: 21 percent high. 47 percent moderate. 32 percent low. Zero percent insufficient information.

CO-CHAIR STILLE: And finally overall suitability. Any last comments?

(No comments.)

CO-CHAIR STILLE: Okay. Let's vote.

MS. ALLEN: Overall suitability for

endorsement for Measure Number 2624, Functional Outcome Assessment. One, yes. Two, no.

Voting starts now.

(Voting.)

MS. ALLEN: 53 percent yes. 47 percent no.

CO-CHAIR STILLE: Okay. So, lots of gray zone things. Anything, Sarah, before we go to member comment?

Okay. Right before we break for lunch, member and public comment is open including the folks behind us.

MS. GHAZINOUR: Operator, would you please open the lines for public comment?

THE OPERATOR: At this time if you would like to make a public comment, please press *1 on your telephone keypad. Again, that's *1 to make a public comment.

(Pause.)

MS. AUTREY: Hello. This is Sophia Autrey with CMS. Can you hear me?

CO-CHAIR STILLE: Go ahead.

MS. AUTREY: Okay. So, I just want to be clear on some of the reasons that were identified that were issues for the reliability and validity.

So, are we -- am I to understand that most of the issues are surrounding the possibility or probability of gaming the measure and that's why there was hesitation, or on the reliability and validity?

CO-CHAIR STILLE: I think gaming was a relatively minor issue. I think a lot of people were wondering about sample size. There were some inter-rater reliability things, if I remember correctly, that were kind of borderline. Different things.

I think that having some new data from 2013 will be helpful. And what does everybody else think generally?

MEMBER BEVANS: I think that greater clarity on how each element of the process definition is actually measured in the field would go a really long way to help us understand

what this measure is actually getting at.

And by extension, whether it is an important outcome or important process.

MS. AUTREY: Okay.

MEMBER KAPLAN: Yeah, I echo that. This is Sherrie Kaplan. I echo what Katherine just said.

I think the link between reporting a functional status measure in one place and a care plan in another place, and they could be for completely different problems as long as they were done, and done is all we're being measured, would help us interpret what the measure is actually getting at, what it means.

So, I think that link was the thing that was most disturbing at least for me.

MS. AUTREY: So, what I'm hearing is the fact that you would want additional information on how the measure is actually operated at the level within -- for the physician, and you want to know specifically the link between the functional status outcome and

the care plan.

Wanting to know that established link documented, or we just have to figure out something that -- I'm just trying to figure out how that would be identified in the measure if you are not trusting that the physician that is putting the information in is clearly putting what has been done.

MEMBER BEVANS: For me, that's less the concern, in part, because it will be very difficult to mitigate that.

It's more understanding what are your coders looking for. What exactly are the criteria that are used to check the box, yes, a measure was used, yes, there is a care plan?

That is linked to outcomes from the functional status assessment.

MEMBER KAPLAN: Yes, I agree. If it was a shoulder pain functional status assessment, one would like to see a shoulder pain care plan or something that --- and if that's not doable, then some clarification about exactly what

documentation we actually are making and then give some feedback on how to improve that performance because, you know, if you're going to use it in quality improvement and public reporting, how do you get a better score if those two things really can't be linked?

MS. AUTREY: Okay. All right. That's clarified. Thank you.

MEMBER CELLA: So, just one more comment. I might put a different spin on it, the same basic idea, because I agree with what's being said.

But what about imaging a case where the numerator isn't just populated by a one or a zero, but that there was some way to, you know, because I keep keying in on the capital AND and the based on bridge in the numerator statement of the indicator.

And in order to achieve that, there has to be some way for somebody to document that they're linked and that it actually did flow that way, but maybe that could be a bonus.

Maybe the current reporting of, yes, there was an assessment done and, yes, there's a care plan gets you one point, but maybe showing the link gets you two points.

So, I don't know if you want to think about it that way, but that would make a little more sense to me because then somebody could be getting a bonus for doing more than reporting on the two components of this linked measure.

MS. AUTREY: Yeah, this is something to think about. I think that one of the issues identified as far as specifically quantifying the operationalization of the measure is really a key point.

So, I appreciate your feedback. Thank you.

CO-CHAIR STILLE: Okay. Any other member or public comments?

THE OPERATOR: And there are no public comments at this time.

CO-CHAIR STILLE: Thank you.

CO-CHAIR PARTRIDGE: We are proceeding

to lunch. If we look at the agenda, the next item up is discussion of 2653 with -- is Minnesota going to be on the line, or are they here?

MS. PITZEN: Hi. This is Collette and Jasmine from Minnesota Community Measurement. And we are on the line ready whenever you guys are.

CO-CHAIR PARTRIDGE: Good. Well, we are just about to decide when that is.

MS. PITZEN: Okay.

CO-CHAIR PARTRIDGE: Quarter to 1:00? All right. 12:45 our time. 11:45 yours in Minnesota. And it's snowing here. I just want you to feel that we feel your misery.

MS. PITZEN: Is that okay if we just stay on the line until you return and we'll just mute ourselves?

CO-CHAIR PARTRIDGE: All right. We will skip the item scheduled at 1:10. And depending on time, we may move one of the -- some of the discussions scheduled for tomorrow

afternoon into that slot around one o'clock.

MEMBER MONROE: Excuse me, Lee. Did we move 422 off the agenda as well?

CO-CHAIR PARTRIDGE: 422 is off. It's in the group. I'm not sure we took a formal vote on 422.

(Whereupon, the above-entitled matter went off the record at 12:23 p.m. and resumed at 12:57 p.m.)

CO-CHAIR PARTRIDGE: I think we are about ready to come back if most everybody is here.

CO-CHAIR STILLE: Pretty much. Mitra asked me to let everyone know that dinner, for those who are interested, tonight is at 6:00 p.m. at Mio, which is right across the street from the Residence Inn, for those of you who are staying there. And Mitra has the address if you need it.

CO-CHAIR PARTRIDGE: Okay. Our next measure is 2653, and our developer is the Minnesota -- I always get this mixed up -- Minnesota Community Measurement.

And, Minnesota, are you back on the line?

MS. PITZEN: Yes, we are.

CO-CHAIR PARTRIDGE: Well, welcome.

MS. PITZEN: Thank you.

CO-CHAIR PARTRIDGE: And would you like to proceed and give us a little description of the background for this measure and what it is you intend that it do?

MS. PITZEN: Great.

Good afternoon, everyone. I'm Collette Pitzen, a measure developer with Minnesota Community Measurement. And with me is Jasmine Larson, our Manager of Measure Development.

We are pleased to be presenting the results of several years of development work for some new patient-reported outcome measures related to postoperative functional status.

The first measure that we are talking about today is number 2653, Average Change in Functional Status Following Total Knee

Replacement Surgery.

This is a measure that is evaluating the change between a patient's preoperative functional status and their knee function one year postoperatively. It is an outcome measure, but its construction is a little bit different from a traditional measure with a numerator or target. Rather, it is assessing the average change in the functional status and has no numerator. I would like to spend a little bit of time walking through some of the measure construct details.

The initial patient population is adult patients age 18 and older, with no upper age limit, who undergo either a primary total knee replacement or revision total knee replacement with dates of procedure during the calendar year.

The measure focus is the orthopedic practice, and procedures are identified using the CPT codes that the surgeons use to bill their professional fees. There are no upfront

exclusions for the initial patient population, and outcomes are stratified by primary or revision procedure type.

The measure is a patient-reported outcome or PRO-based measure. The PRO tool that is used is the Oxford Knee Score tool, or OKS, a 12 question tool selected by the Measure Development Work Group for its strong psychometric properties, easy for patients to complete, and simplicity in administration and scoring.

The patient completes the OKS anytime within three months prior to the date of the procedure. The patient then completes a full postoperative assessment at one year, with a fairly wide window to capture as many postoperative assessments as possible. One year is defined as nine to fifteen months postoperatively.

Change is first calculated for each patient, and then the changed scores are summed and the average is determined. The measure

calculation takes into account both patients that have an improvement and those patients whose function decreases postoperatively. In order to calculate the change in each patient's functional status, the measure denominator is comprised of patients who have a completed preoperative and postoperative assessment.

It is important to understand the rate of tool administration in the population prior to any use or reporting of outcome measures, and we accomplish this through paired process measures and the submission of all patients for rate calculation, even those patients who may be missing a PRO assessment. Paired process measures and the inclusion of all patients is one way to address potential gaming of this measure.

The first-year results for the measure demonstrate a 17-point increase on a 0-48 point scale, where a higher score indicates improved knee function. Variation is noticed based on the annual volume of TKR procedures performed, with groups performing 100 or more procedures per year

having a higher average functional status change.

Thank you for the opportunity to present this measure for your consideration, and we welcome your discussion and questions.

CO-CHAIR STILLE: Collette, I think it would be helpful to the Committee if you would also give us a little background as to why Minnesota Community Measurement undertook to develop this measure. What it was that --- I know it comes up again in the other measure we have before us today, but I think most of the Committee is probably not familiar with the process in Minnesota of how you develop measures, why you develop measures, and how they are now used across the State.

MS. PITZEN: Sure, I would be happy to.

This measure -- we are a subcontractor to the Minnesota Department of Health. As part of that subcontractor relationship, we also work on developing new measures in addition to publicly reporting and using our measures in a

statewide quality reporting and measurement system.

And we are frequently presented with a concept for measurement development for exploration of determining can excellent measures be built around a particular topic. This topic of total knee replacement was presented to us in 2010.

So, when I talk about a couple of years of development, we have been working on this for a while. Part of the rationale for the selection of total knee patients is the anticipated large boom in volume of procedures over the next, I want to say, 10 years or so, with the Baby Boom population.

Part of the reason about the length of development time, because we are looking at a postoperative assessment period of 15 months, it did take us quite a bit of time to complete testing of this measure. Thank you.

CO-CHAIR PARTRIDGE: As I understood the narrative, one, the Department of Health in

Minnesota is concerned in part about the potential for overuse here. Is that correct? You see a boom. You want to be sure, given the dollar volume and the number of procedures, that your money is being properly spent. Am I putting words in your mouth?

MS. PITZEN: Just a tiny bit. Perhaps not a concern with overuse yet, but I think there is some underlying currents of that. But, rather, it was having really a lack of information for consumers to know what they could expect after undergoing this procedure.

And frequently, our work with the orthopedic and neurosurgeon groups and other specialties, oftentimes it's anecdotal. And so, this is a new effort to try to quantify and put some information together about what the outcomes are for this patient population and the spine measure that we will be presenting next.

CO-CHAIR PARTRIDGE: Good. Thank you. Dawn?

MEMBER DOWDING: Okay, thank you. In

terms of --- do you want me to just talk about importance first? Because I, then, have some issues about reliability and validity, but that comes later.

I think in terms of the description of importance, you have made a very good case for why at a national level we might be concerned about total knee replacements in terms of variation.

But I just wondered if you had any data from your pilot study in Minnesota, apart from size of practice, to illustrate variations in performance gaps for other factors such as age, ethnicity, and how different stratifications of patients, how the average difference may appear.

And one of the other things -- it is just a very general comment -- with a lot of these scores is that what we are actually asking --- being asked to endorse is the difference between preoperatively and postoperatively and how big that difference is, and whether that is

meaningful.

And I am not entirely sure the difference of 14 points to 17 points is actually meaningful, useful, different for the patients in terms of function on the scale. I just don't have any feel for what it actually means in terms of quality of care.

So, I just wondered if you could just talk us through some of those issues to do with this. Is there actually a gap in quality of care associated with these knee replacements, and how would we know that from this different score?

MS. PITZEN: This is Collette. Thank you very much for the discussion and questions. I am not entirely sure that I have all of the answers. I know that we --- and we struggled with performing some of the reliability statistics. In many of our other measures we can demonstrate meaningful differences between the practices and opportunities for improvement.

This is a newer-type measure. We believe that, as we keep going forward, that we

will be able to, like several other measures, have discernible differences between the practices.

Just to share a little bit about the Measure Development Work Group's thoughts around this patient population, though we are not specifically measuring, they felt that there could be some differences. We also have a three-month assessment measure that we did not put forward for endorsement, but the thought and feeling at the time was a three-month assessment of the patient, while not reflecting their full function, could discern differences in postoperative rehabilitation and perhaps surgical techniques used and the selection of patients.

So, there was that consideration as we went forward. And the Work Group felt that, with this brand-new measure, that differences would be demonstrated among the practices.

CO-CHAIR PARTRIDGE: Questions from other members of the Committee about importance?

Sherrie?

MEMBER KAPLAN: Thank you. I guess I have a followup question on that. What is the effect size of that difference? What proportion of standard deviation is it?

MS. PITZEN: This is Collette. I missed the initial, the first question, but we don't have that information in terms of the standard deviation.

MS. LARSON: We don't have it in front of us right this moment.

MEMBER KAPLAN: If you could give us a sense, that would help us interpret those differences you are observing as meaningful.

MS. PITZEN: I don't know how much that we can pull together on the fly here. We will give it a try as the discussion keeps going.

CO-CHAIR PARTRIDGE: Collette, I am looking at my notes from this measure, and I notice on page 16 of your attachment you did note that was a variation in the performance among the four regions of Minnesota on this measure, and that the range was from 10.8 to 12.9.

And I think the question Dawn asked is similar to the one that I have. We would be curious to know that if you think that -- and I should say some of my colleagues on this Committee could probably answer this question for me, too. Is the two-point difference a significant one?

CO-CHAIR STILLE: It all depends on the distribution and the number of patients and stuff. It depends a lot on the number of patients and the distribution of the scores and that kind of thing.

CO-CHAIR PARTRIDGE: Go ahead, Sherrie.

MEMBER KAPLAN: Yes, if we don't have the standard deviation, we can't interpret the effect size --

CO-CHAIR STILLE: Right, right.

MEMBER KAPLAN: -- and the magnitude of how much that difference is.

CO-CHAIR STILLE: Right.

CO-CHAIR PARTRIDGE: Okay. And part

of this is we are dealing with a pilot. So, we have a small sample.

MS. PITZEN: That's correct.

MEMBER BRADLEY: In the study, did you also collect data on interventions postoperatively that might have affected outcome? Say, patients that had rehab versus patients that did not have rehab?

MS. PITZEN: This is Collette. No, that we did not collect that information.

I just wanted to step back a second and describe our standard processes for the measures that we collect and report in Minnesota. And these are statewide measures.

We have a philosophy of colleting minimal datasets for what is necessary for risk adjustment and calculation of the measure. So, as we are working through our development process, we really caution our work groups, in our measure design and construction and the actual data fields that we are requiring groups to submit to us, to really be mindful towards

burden and not be asking for every possible element that they can think of.

MEMBER DOWDING: Yes, and just a followup question as well. I was also wondering why you decided to follow up a year post-surgery, with that sort of time scale, because it just seems to be quite a long time after the surgical intervention, and what effect that has on the state's rates.

MEMBER BIERNER: Well, let me just comment that that is actually, I think, a good thing. Joint replacement operations, patients often may have symptoms and don't really -- I think six months would be the minimum I would want to see anything after a joint replacement because there is a significant recovery time, depending on the age of the patient. So, I think longer is actually better, and six months to a year is probably appropriate for a joint replacement.

MS. PITZEN: Great. This is Collette. Can I just add an additional comment?

We did have a really thorough discussion at our work group level about the timeframes that they wanted to assess. I initially was suggesting a six-month timeframe, and the orthopedic surgeons very quickly shared those exact same feelings, that the one-year postoperative assessment was really hitting the patients at their true level of functional status improvement, and that to measure much sooner would not do justice to the measure. So, we do have that very long followup time.

MEMBER KAPLAN: Can I just follow that up real quickly? This light test/retest reliability, where if you don't choose the right interval, true scores can vary because you measure somebody at one point, and then, they walk in front of a bus. And you measure them three weeks later, and their health looks really different.

But this is the exact same problem with attribution. If you let the interval go too long -- and this is a content thing, so I am

talking way out of my depth here -- but if you let it go too long, there are lots of things that happen around the quality of care in between or that can happen to really move that score, rather than the attribution.

So, what is the attribution? What is the source of the attribution on this measure? Is it the baseline surgical procedure or?

MS. PITZEN: This is Collette. Is it okay if I jump in when I think the question is being directed to me?

CO-CHAIR PARTRIDGE: Yes.

MS. PITZEN: The attribution is to the surgeon who performed the procedure and thereby his practice.

MEMBER LOEB: Back to when someone had asked, you know, are you basing this on whether there was therapy or something, and what Lee had said is, are we using this ultimately to look at overuse?

As someone who took care of post-op total knee replacements, I think if this

ultimately is going to be used for someone looking to see, is there improvement by having this done, should I have this done, I think it is really important to know whether or not someone had therapy, because there is a huge difference. I mean, if you don't go for therapy after your knee replacement, you are probably worse off than before your knee replacement. So, I think that is one thing that is really missing.

So, this is definitely just a really broad, generalized -- and I am not sure how much 17 points on a scale of 48, if that shows much improvement. You know, going through a major surgery like that, I would want a lot more improvement. That's not even 50 percent. Just a thought.

MEMBER BIERNER: May I say something? I am just looking online. There is a large study of over 3,000 patients using the same Oxford Knee Score, and the standard deviation for the mean score was 8. So, you are talking about a two-standard-deviation difference. And so, that is

potentially very significant, a 16-point difference. And they were looking at the difference among surgeons and how many operations per year each surgeon performed who does this operation.

So, I think this is measure is useful because the standard of care is knee arthroplasty for persons with end-stage degenerative arthritis of the knee. And so, there are not a lot of other treatment options one is going to be looking at. One really looks at is how well done was your knee surgery versus someone else and how good your functional outcome is at one year. In this study of 3,000 patients they used six months and two years. So, it is not unreasonable to use a one-year point.

CO-CHAIR PARTRIDGE: Becky? Peter?

MEMBER THOMAS: I am not a clinical person, but I do know a fair amount about joint replacements and policy around them and rehab potential. And so, Sherrie's comment about rehab certainly struck a chord with me.

But the question I have is under threats to validity in terms of the risk adjustment. I am also aware that many joint replacement patients have no comorbid conditions and some have major multiple complications and comorbidities. I am just wondering, there are no exclusions in the denominator and I am just trying to search for how you might have accommodated that in the risk-adjustment methodology.

MS. PITZEN: This is Collette from Community Measurement.

I can appreciate that. There is a couple of different things for consideration.

Originally, the Work Group talked about upfront exclusions, keeping in mind the overall incidence of all those, particularly like a typical exclusion is death, a very small percentage in this population. And the actually denominator for the measure are those patients that had a pre- and post-operative assessment. So, patients who died during the assessment

period are not included in the measure. So, that was part of their thinking about initially no upfront exclusions.

But we have several data elements that, when we have more data, when we have more patients, we plan to run these data through our risk-adjustment models, and those elements are included in the documentation. I will share them with you now. Age, gender, zip, race/ethnicity, country of origin, primary language, insurance product as a proxy for socioeconomic status.

In addition, part of our Development Work Group discussion is other clinical variables that are felt to be important. For this population, we also plan to include the pre-operative functional status Oxford Knee Score. We are actually collecting some quality-of-life data that is not part of this particular measure construct, but can be used in quality of life. So, we have those pre-operative scores.

We are also looking at the patients' BMI, the diagnosis of diabetes, and tobacco use

of the patient. So, our plan is to put all of those variables through our risk-adjustment model to determine which are relevant variables to use going forward.

Oh, and I just wanted to mention, too, the denominator does include primary knee replacement and revision knee replacement. We are collecting and reporting data on both of those procedure types, but they are not reported together. They are reported separately.

MEMBER THOMAS: So, a quick followup. For the physicians or clinical people in the room, is that sufficient to risk-adjust those with major multiple comorbidities or complications or not?

MEMBER BIERNER: Yes, I think so. Because Body Mass Index is a known negative predictor. The surgical deaths after total knee arthroplasty are going to be related a lot to those comorbidities. And so, they are not captured --- you know, those who die during the period would not be included in this study. But

I think those are useful ones.

CO-CHAIR PARTRIDGE: I don't want to cut off discussion if it on importance. If it is related to reliability and validity, I would like to hold.

So, Carin, Dawn, and Ann. Yes? Go ahead.

MEMBER MONROE: If I am understanding you, this is a measure of a surgeon about how well his folks are doing a year later. Is that correct?

MS. PITZEN: That is correct.

MEMBER MONROE: But you are not looking at anything that happened between the surgery and the measure a year later. Is that correct?

MS. PITZEN: This is Collette. Let me clarify.

So, you are asking me if we are collecting process measures associated with particular things that were done to the patient. And my answer to that is no. But I also wanted

to qualify it with a philosophy that we have in Minnesota. As we go forward with our transparency and public reporting, we start getting into identifying best practices among the different participants in the measure.

I frequently get calls on our depression measure; you know, what is Mayo Clinic doing? Their rates are fabulous. We would like to connect with someone there to understand how they are achieving excellence. So, we do have that philosophy.

MEMBER MONROE: Well, I would appreciate that, but I would think you are losing some opportunity here with this measure to really identify what might have made a difference. Because to hold a surgeon accountable -- I have had two knee replacements in the last year -- and if I hold my surgeon accountable for what I am doing a year from then, that's a pretty long period of time, whether I had rehab or not, whether I followed instructions, whether I even had rehab at home or went to a facility.

To me, there is so much that could happen in that year, that to hold a surgeon accountable for that seems both a missed opportunity and perhaps a misassignment of responsibility. That's just my comment.

CO-CHAIR PARTRIDGE: Any other questions on importance before we vote? Deb?

MEMBER SALIBA: So, I do want to speak to the idea, however, that surgeons should be held accountable for outcomes beyond hospital discharge. I mean, my husband had a knee replacement this year, too. And I would be telling the surgeon, "Well, aren't you going to send him for rehab?" when he was sending him home.

And I think there is a patient activation model that really involves the physician being part of that activity. So, one year may be too long, but I think there is something to be said for starting to think less about just the surgical episode and more about that surgeon is responsible for interacting with

the healthcare team. If rehab is not going well, figuring out why rehab is not going well, and sort of more than just being the procedural list.

If they are sort of the leader of that team of persons taking care of that knee replacement, which the way the system is set up right now, they are the lead of that team to a large extent. I mean, they are the ones that are interacting with rehab and with the physical therapists and are not in most cases.

But, anyway, I just want to put in a plug, as you think about how to modify this measure, that we really do want to start to encourage more long-term outcomes. Maybe a year is too long, but, yes.

MEMBER LOEB: I am just going to jump in really quick. I think you said the perfect word in "team." I mean, this is not just a surgeon. This is not just the patient. This is a team. And a surgeon can't have dynamic outcomes without the patient working with him, and the patient can't have dynamic outcomes

without a good surgeon. So, it needs to be a team with every single one of these measures that are measuring the outcome of the procedure. So, I mean, that word's vital.

MEMBER MONROE: I have just one other point.

CO-CHAIR PARTRIDGE: Ann, go ahead.

MEMBER MONROE: Were patients involved in looking at whether or not this was a useful measure for them?

MS. PITZEN: Yes, they were. They are part of our Measure Development Work Group.

CO-CHAIR PARTRIDGE: Thank you.

All right. Are we ready? Nadine?

MS. ALLEN: We are voting on evidence. The rationale supports the relationship of the health outcome to at least one of healthcare structure, process, intervention, or service.

One, yes; two, no. The voting starts now.

(Voting.)

All votes are in. Eighty-four

percent, yes; 16 percent no.

Voting on performance gap. One, high; two, moderate; three, low; four, insufficient.

(Voting.)

All votes are in. Sixteen percent, high; 63 percent, moderate; 16 percent, low; 5 percent, insufficient.

Voting on high priority. One, high; two, moderate; three, low; four, insufficient. Voting starts now.

(Voting.)

All votes are in. Forty-two percent, high; 47 percent, moderate; 11 percent, low; zero percent, insufficient.

CO-CHAIR PARTRIDGE: On to reliability and validity.

MEMBER DOWDING: Okay. Correct me if I am wrong, but the data that you have provided for the reliability and validity of this measure is actually taken from the original study that developed the Oxford Knee Scale in 1998 that looked at the reliability and validity of the

scale.

And we don't actually have any data from your pilot study to show that the measure, which is the difference between pre-op and post-op one year is reliable and valid in its use in Minnesota. Is that correct or have I read it wrong?

I have been through this form about four times trying to find some description of the patients that it was used on, some response rates, how many people actually filled in the form pre- or post-op one year, what their makeup is in terms of population, what age they are, gender, ethnicity.

And I can't find any data at all on that, and I can't find any data related to the pilot study to do with reliability and validity, just the original scale development data from 1998. And being British, I can guarantee it was probably British patients in Oxford that filled it out. And I am not necessarily sure they are the same as people in Minnesota.

MS. LARSON: And so, this is Jasmine at Minnesota Community Measurement.

When it comes to reliability testing of the performance score measure itself, the nature of the problem is that it is a new type of measure, and there isn't a traditional numerator/denominator. And it is a continuous measure based on an eligible patient population.

And we worked with NQF staff to try to identify the appropriate testing for reliability at the performance score level. To our knowledge, the appropriate methodology to be applied in this type of scenario has not been established or performed by NQF or other measure developers of measures of this nature. So we don't know that the measurement science has evolved to the point of determining the appropriate methodology for testing reliability at the performance score level.

MS. SAMPSEL: And so, this is Sarah, and I just want to kind of talk through exactly how we have been working with Minnesota Community

Measurement as well as other developers that will be presenting over these two days.

We had identified, with Jasmine and Collette a couple of weeks ago that, while we were doing the measure reviews, that the measure level reliability scores were not there. They identified that as well.

And we were looking at kind of other potential examples within the NQF portfolio, and we found one, but we didn't share that back with the developers because we wanted to have this conversation with you all first, and kind of how we did with the last time, once we have some clear direction to the developers, give them the opportunity to provide that data.

And I don't know, Sherrie, if you have other examples. But the one that we found was actually some testing that Yale had done. I don't think it was an admission measure. I think it was a mortality measure. But they had done some testing on interclass correlations. We thought that might be a good fit, but we weren't

sure, which is why we wanted to make sure this Committee had the discussion that we can give them clear direction on what to do.

CO-CHAIR PARTRIDGE: Sherrie?

MEMBER KAPLAN: Yes. First of all, there are ways. And usually, what we do is the interclass correlations will tell you how much between subject reliability versus within subject reliability there is. If you are going up to the physician level, you are dragging along the patient-level errors of measurement and you are creating a composite, then, at the physician level.

What you would then do is look at interclass correlation coefficients for between versus within physician-level variability. And there are ways to accommodate both errors, and there is a thing called the Spearman-Brown prophecy formula. It sounds really spooky, but it is not. It is how many measures at the patient level do you need to aggregate and create a composite at the physician level that will give

you a certain level of precision.

So, there are certainly ways to do this. And interclass correlation coefficients would be the most reasonable in this circumstance. That is what we asked of the earlier measures developers as well, to give us a sense of the interclass correlation coefficients and the magnitudes.

So, can you discriminate and, if not, you certainly can't, almost never, discriminate one physician from another. But are you trying to discriminate the tails of the distribution from each other? Are you trying to discriminate a benchmark? It kind of depends on what the measure is ultimately going to be used for.

MEMBER DOWDING: And I also think just in general just some idea of the patient population that this was piloted on would be helpful and some indication of what the response rate was in terms of pre- and post-operatively. Is it 50 percent? Is it 70 percent? How representative of the patients? If they are

filling it in themselves, is it in other languages? Disparity information? Just some idea of the sample on which it was tested on.

We have some idea about the practices and how many operations they did, but no insight into the patients who were in the piloting. And that would be very helpful information.

MS. PITZEN: Great. This is Collette at Community Measurement.

We actually had two phases of pilots, so my apology. I have some demographics on the first phase of our primary knee replacement patients, and I would be happy to share that with you.

The first phase of the pilot had 1100 patients in them. The majority of those were primary knee replacements, 92 percent. Our population was a little bit higher on the female side, with 59 percent female, 41 percent male. The average age was 64.7 years, with an age range between 36 and 93 years.

We also are collecting the location

where the procedure is happening. We have some movement to total knee replacement in the ambulatory care setting as well, and that reflected about 18 percent of our patient population.

In terms of race/ethnicity, we had fairly good capture of that data element from our pilot participants. About 91 percent were able to report race/ethnicity, and 95 percent of the patients were White.

Not as great capture in country of origin. However, primary language was captured almost 100 percent. And again, we have a very high English-speaking population.

I have some additional statistics around the risk-adjustment variables. Tobacco-free was about 87 percent. So, that would be about 23 percent smokers -- or, I'm sorry, 10 percent smokers.

We have a 14-percent incidence of diabetes mellitus in the pilot participants.

And for this particular measure -- and

I want to qualify difference than our lumbar spine measure experience -- we did have some difficulty with the PRO tool administration rates. And we are choosing a phased approach to try to get those PRO tool administration rates up, so that we are able to report the outcomes.

CO-CHAIR PARTRIDGE: Are we ready to vote?

MEMBER KAPLAN: Can I ask one more question from the NQF staff? And this is not necessarily for this measure, but for all measures. Is it not reasonable to at least ask at the patient level what the standard error of measurement is? Standard error of measurement is the standard deviation times 1, minus the reliability.

So, with a very reliable measure, you get almost the standard deviation. With a less reliable measure, you get it amplified. So, distinguishing scores is going to be helped if the measures developers can provide that information. And certainly, even better, the

unit of inference that we are trying to draw, like the surgeon versus the patient versus the clinic.

MS. SAMPSEL: Yes, I mean, that is definitely reasonable, and I think there are times when -- I don't know if we err where we go on the side of caution to some degree, where we throw out some possible things and datapoints that could be provided to us.

And I think some of the measures that we have encountered in this project are helping us learn, too. So that I think that we could be more prescriptive in the future from this went forward with the measure developers for these types of measures, yes.

MS. LARSON: This is Jasmine at Minnesota Community Measurement, if I could just comment that I don't know that this is necessarily what was asked previously regarding the standard deviation and the effect size, but I was able to calculate the standard deviation at the reported entity level of 4.22 for this

measure. But it is not at the patient level, but 4.22 for average change at the medical group level.

CO-CHAIR PARTRIDGE: Okay. Ready to vote?

MS. ALLEN: Voting on reliability.

One, high; two, moderate; three, low; four, insufficient. The voting starts now.

(Voting.)

All votes are in. Zero percent, high; 47 percent, moderate; 16 percent, low; 37 percent, insufficient.

Voting on validity.

One, high; two, moderate; three, low; four, insufficient information. The voting starts now.

(Voting.)

All votes are in. Five percent, high; 37 percent, moderate; 26 percent, low; 32 percent, insufficient information.

CO-CHAIR PARTRIDGE: Discussion on feasibility.

MEMBER DOWDING: I wonder if you could just clarify for us. I think you have alluded to it a couple of times, that you have actually had difficulty with this measure, getting patients to fill it in and send it back. Am I right? I think you have mentioned that a couple of times.

So, could you just talk us through how it is administered and how you would treat the data, so we can get some idea of how feasible it would be to collect routinely.

MS. PITZEN: Sure. This is Collette again.

As part of our pilot-testing process, we are doing full measure specifications and, also, field data element specifications for the information that is needed to calculate the measures.

For this particular measure, we are asking for a pre-operative OKS summary score. The tool is simple to sum and score. There is not a complicated algorithm or formula that needs to be applied.

We have a process in Minnesota called direct data submission where the practices submit through a HIPAA-secured data portal, patient-level information that is needed to calculate these measures. And the data dictionary, which was quite extensive, was provided.

During our pilot process testing, we also are constantly working with the group in terms of questions and answers, but when we are completing the pilot, we are serving them for burden. And actually, measurement is new to the orthopedic practices. Unlike primary care, which has been accustomed to measuring and collecting data for quite some time, it is new in the orthopedic world.

There are a couple of different EMR systems that these practices use. And groups were successful in building retrievable or discrete fields within their EMR to capture this information and extract it back out.

And in fact, the groups, when they were rating things that were challenging for

them, rated the actual EMR build and store of this information as less challenging than getting the patient-report tools into their workflows.

So, if I could just briefly jump ahead a little bit to the lumbar spine measure, the Oxford Knee Score tool was used by a couple of groups in our State. Nobody was routinely using anything. So, it was a newer tool to be implemented into clinical workflows. And we did see that with lower than we would like pre-operative administrative rates of the Oxford Knee.

You have a captive population. You would hope that you would be building that into your pre-operative paperwork and workflow process to administer that tool as you are assessing the patient and planning for their surgery.

So, our pre-operative rates on average were less than 40 percent right. Some of them, some groups were performing at a very high rate, but on average it was about 40 percent.

And then, as they gained acceptance

and ease and familiarity with this process, we actually saw many groups that had a lower preoperative Oxford Knee Score rate, actually, their postoperative rate was higher than their preop. So, we are seeing this gradual implementation of these tools into the clinical workflow process.

So, our Measurement and Reporting Committee, which is our approval body for everything that we are publicly reporting, approved a plan to publicly report the process measures that go along with this for tool administration. And we have actually published that on Minnesota Health Scores. So, we are hoping to be in a place where we can publish the outcome scores in the next submission year.

Does that help answer your questions?

MEMBER DOWDING: Yes, for the pre-op, but post-operatively how were they filled out and what is the completion rate in terms of the percentage of patients who have both measures?

MS. PITZEN: Sure, sure. And in part

of our Work Group development process was the surgeons on the Work Group were seeing about 70 percent of their patients routinely at one year. So, we are were kind of shooting or hoping. We would never expect 100 percent post-operative capture of patients.

So, that the data can be collected in a variety of different ways. We had several groups that would, if they weren't seeing the patient in clinic in that timeframe, they would mail out the questionnaire to the patient and have that returned or pushed out by their patient portal or EMR, and have that returned.

So, again, on average, for this -- again, we are pilot testing -- for this measure, on average, we had a post-op rate of around 31 percent, but some variability, with some groups achieving close to that 70 percent mark post-operatively.

So, because our pre-op rates were so low, the rate of having the denominator of pre- and post-operative gave us information that we

couldn't work with right away.

CO-CHAIR PARTRIDGE: Any further comments, questions?

Are you ready to vote on feasibility?

MS. ALLEN: Voting on feasibility.

One, high; two, moderate; three, low; four, insufficient.

Voting starts now.

(Voting.)

Please point your clicker towards me. Please try again.

It is probably frozen. One second. Experiencing some technical difficulty.

CO-CHAIR PARTRIDGE: Collette and Jasmine, our vote register isn't working at the moment. That is the silence.

MS. LARSON: Okay. Well, Collette, you know, is a member of the Surgery Committee Standing Committee. So, I think she is familiar with the technical hiccups that can happen during these meetings.

CO-CHAIR PARTRIDGE: Yes. Go ahead.

MEMBER BRADLEY: I guess it just kind of occurred to me that, if we are really measuring the physicians -- and there was some reference to the physician does some screening prior to surgery to assess for appropriateness of the patient for surgery. So, there is some kind of risk adjustment that the surgeon does on the front-end that is not reflected here.

But was there ever any discussion that, because this is physician-reported outcomes, that they may, again, game the system by selecting patients that have the highest potential for outcomes, and then, patients who need this surgery, but perhaps have comorbidities or have other issues may not be eligible for surgeries from some physicians because they are trying to get their scores up?

MS. PITZEN: This is Collette. I would be happy to answer that question.

Actually, our Measure Development Work Group did also explore developing appropriateness criteria. And currently, there are no national

society guidelines for appropriateness criteria for knee replacement. The guidelines actually intentionally stop at that point. We had that discussion and we could not come to consensus or a resolution about appropriateness as a measure.

But I just wanted to share that we are collecting that preoperative functional status score, and that there is a plan to evaluate that for risk adjustment. So, we are not setting -- you have to at least an OKS score of such to be in the measure. We are taking all patients.

In terms of gaming, again, I want to explain our process here as well. We are -- and it is state mandated by law -- we are collecting the data on all patients. Regardless of if they had an assessment or both assessments completed, we are collecting the information on all patients.

And if there is a low percentage of tool administration, one, we can't reliably report outcome measures, but that was one way to address potential gaming.

CO-CHAIR PARTRIDGE: Thank you.

I think we are going to switch to low-tech, so we can complete this measure.

So, let's use a hand vote, beginning on feasibility.

And in favor of high?

(Show of hands)

Okay. Moderate?

(Show of hands)

MS. THEBERGE: All right. Keep them up for a minute longer.

MS. ALLEN: Okay. Low?

(Show of hands)

And insufficient?

(Show of hands)

MS. THEBERGE: One high; 15 moderate; two low, and one insufficient.

CO-CHAIR PARTRIDGE: Any other questions?

CO-CHAIR STILLE: I guess just going back to the spread in differences between groups, which we were wondering what the effect size was

and whether that was clinically meaningful. And then, I think the measure developer said that the standard deviation was about four.

So, if that is the case, you know, we have got the difference between 14 point something and 17 point something, and the standard deviation of four. I wonder how usable it will be to discriminate between groups, if I have my numbers right and if that is my interpretation of what their numbers were.

CO-CHAIR PARTRIDGE: Collette, do you want to respond to whether or not -- whether Chris has understood you correctly?

MS. LARSON: This is Jasmine actually. And yes, it was -- the standard deviation was around four. And his point regarding the difference between 14 and change and 17 and change is well-taken.

However, there is more spread in the actual medical group that fall outside of the standard deviation range that would allow for classification. I am trying to pull that up

right now. So, just bear with me.

MEMBER KAPLAN: Can I ask a question? I understood that there was no estimation of reliability at the physician level. Is that correct? Right now, they don't have anything for us to put in, so we can't understand what the magnitude of the standard error of measurement is. So, even if we knew the standard deviation is 4.22, for example, you would multiply that times the square root of the difference between one minus the reliability.

So, if we have low reliability, then we get a bigger spread, and it is less interpretative. Then you are pushing the extremes before you get meaningful differences. If we have high reliability at the physician level then we get four. So, the smallest difference we are looking for here is four for the standard error of measurement. So, two-point differences are probably within the standard error of measurement, which would be a little bit disconcerting.

MS. LARSON: Are you looking for me to respond? This is Community Measurement.

CO-CHAIR PARTRIDGE: Jasmine, yes, if you would like.

MS. LARSON: Yes. So, I again take that point well. And I hope, with the additional information that has been shared in this call, that we will be provided the opportunity to run the methods that were described here, because I am confident that we have the information to be able to do that and provide that additional detail for consideration.

CO-CHAIR PARTRIDGE: Jasmine, I think that would be very helpful to all of us, and there is a little time. I think NQF staff will get back and talk with you about that in more detail.

So, vote on usability.

MS. ALLEN: Voting on usability.

One, high; two, moderate; three, low; four, insufficient information. The voting starts now.

(Voting)

All votes are in. Zero percent high; 63 percent, moderate; 26 percent, low; 11 percent, insufficient information.

Voting on overall suitability for endorsement of Measure 2653, average change in functional status following total knee replacement surgery.

The voting starts now. One, yes; two, no.

(Voting)

All votes are in. Fifty-eight percent, yes; 42 percent, no.

MEMBER KAPLAN: I have a question for the NQF staff. If this isn't provisional, there should be a third category, which is no pending results or something like that, or, yes, pending results. Is there no opportunity -- this dichotomy is making me feel real uncomfortable.

MS. SAMPSEL: So, what will happen is -- there are a number of criteria here that fell in the gray zone. So, we would work with

Minnesota Community Measurement staff to have the opportunity to bring back information before the end of public comment. And so, you will reconsider the measure at your post-public-comment call and also be able to consider any public comment that comes in, and you will revote.

CO-CHAIR PARTRIDGE: Okay. Thank you very much, Minnesota team.

MS. LARSON: Thank you.

CO-CHAIR PARTRIDGE: I am just looking ahead, and I see it is the next item on the agenda because the intermediate item dropped out.

So, should we proceed to 2643 while you are on the line? And let's see who has the lead. It is Dawn and Sherrie. Okay, whichever one of you wants to lead off.

MEMBER DOWDING: Okay, it looks like it is me again.

And basically, I have exactly the same comments as for the last measure for this measure. Do you want the developer to talk

first?

CO-CHAIR PARTRIDGE: I apologize, I mixed up.

Liz is the other on this one.

MEMBER MORT: I will be happy to give it a whirl to start. Oh, was Minnesota going to say something first about this measure?

MS. SAMPSEL: Collette and Jasmine, did you have anything you wanted to say to introduce this one?

MS. PITZEN: Yes, please. This is Collette and Jasmine again.

Our second measure is 2643, the Average Change in Functional Status Following Lumbar Spine Fusion Surgery.

In terms of measure construct, there are many similarities to the total knee measure. In fact, both of these Work Groups started their development work at the same time and actually came to some of the same measurement decisions.

This is a PRO-based outcome measure evaluating the change between a patient's

preoperative functional status and their postoperational functional status at one year. The initial patient population is adult patients age 18 and older with no upper age limit who undergo a lumbar fusion procedure at any level or number of levels, including the lumbar vertebrate, during the calendar year.

The measure focuses orthopedic and neurosurgery practices, and the procedures are identified using CPT codes. Exclusions for this measure are cancer, fracture, and infection related to the spine and idiopathic or congenital scoliosis.

The PRO tool that is used is the Oswestry Disability Index, Version 2.1a, a 10-question tool that quantifies functional ability related to low back pain. The Oswestry Disability Index, otherwise known as the ODI, is used widely in clinical practice and research, has strong psychometric properties, and is considered the gold standard for assessing low back pain. The tool is scored to reflect a

percent disability where a higher percent indicates more impairment in function.

We like to remain aligned across measures where it makes sense clinically to do so. And this measure construct aligns with the total knee measure presented previously.

Assessments are completely preoperatively any time within three months prior to the procedure and, again, postoperatively at one year, defined as nine to 15 months.

Because the measure is one of change between pre- and postop functions, the measure denominator is comprised of patients who have completed a pre- and postoperative ODI tool.

Again, in an effort to reduce gaming, the initial population that is submitted for calculation of outcomes and paired process measure includes all patients, regardless of if assessment tools are completed. Again, there is no numerator or target ODI score, and change is calculated as in the previous measure.

Interesting discussions during the

Measure Development Work Group: design and specification of this measure. Lumbar surgery is an effective procedure for many spine conditions, but may be controversial and less successful for some patients, particularly those with degenerative disk disease whose pain may not be originating from the disk.

Originally, the Work Group wanted to focus on one-level fusion for spondylolisthesis only where fusion is an appropriate procedure and patients do well. But, as there is a very narrow percentage of patients who have lumbar fusion procedures, the Work Group evolved to expanding the denominator to be more inclusive and are also collecting the condition for which the procedure is being performed in one of four categories: degenerative disk disease, disk herniation, spinal stenosis, and spondylolisthesis. These categories may be used for further analysis and/or included in variables in the risk-adjustment model.

Pilot results for the measure

demonstrate an average improvement in function of 17.2 points on a 100-point scale and variability in results among the practices.

And thank you very much.

CO-CHAIR PARTRIDGE: Yes.

MEMBER MORT: Thank you very much for the summary.

I want to applaud you for tackling this controversial area in utilization of surgical procedures. In the writeup you pointed out, although you didn't put this in your comments just now, that there is a 15-fold increase in the number of complex fusion procedures performed for Medicare beneficiaries.

So, this is a highly variable procedure. It is on the rise. As you pointed out, as you politely implied, there is a lot of criticism currently of this procedure being done. I can understand why your Work Group wanted to focus it on the narrow indication of spondylolisthesis because the fusion makes a lot of sense mechanically. For those who don't know,

it just means that there is a lateral translation of one vertebrate over the other and potentially it is unstable or could cause damage. So, I understand exactly what you are trying to do and I applaud it in terms of the importance.

Let me stop there and see if there are other comments from other folks.

(No response)

So, the question about gap coming under importance, what I would say is that there is a gap in something here. I am not sure it is a gap in performance as much as it is a gap in the quality of care being delivered. And this might be a tool and a process whereby we could have a better understanding of what patients and what indications actually benefit from this surgery, which is a big procedure. This is not a little thing. This is a big procedure.

So, in terms of the gap, I think there is obviously variation. So, you could imply that there is a gap in quality of care based on that. I am less clear about the gap in terms of

variability in performance based on the pilot data.

Can you say a little bit more about the variability in performance and your understanding about that element of importance?

MS. PITZEN: Sure. This is Collette.

This may sounds like an apology, but this measure has gone through one phase of pilot testing. We had the opportunity to bring it forward to NQF during this project.

We had four practices that were participating in the pilot. Again, this is in the Statewide Quality Reporting and Measurement System for Minnesota, required of practitioners.

We are expecting full implementation data to be coming in this spring. So, we will fairly quickly have much more data than we have available to us today.

I am just going to share one more thing. Unlike the total knee measure pilots that we did, there was a much higher rate of tool administration in the pilot participants. As

indicated earlier, these patients have been under intense scrutiny from health plans for prior authorization. The Oswestry Disability Index tool is oftentimes required to be administered and submitted to health plan preoperatively to confirm a need for the procedure.

So, we have administration rates that are approaching our desired standard in terms of having pre- and postoperative assessments. So, we do only have four practices that we were comparing the variability between.

MEMBER MORT: So, you expect more information along these lines as the pilots mature?

MS. PITZEN: Actually, full implementation of all practices in Minnesota. The data will be coming in in May.

MEMBER MORT: I would say that, in terms of importance and priority, getting a better handle on whether or not we are subjecting -- I don't mean to say it that way -- whether or not we are offering procedures to patients, a big

procedure for patients who may not do well at all, I think it is obviously a high priority. And this is more than just an outcome measure, as I see it. I think it is a tool to actually change practice, which is beyond the scope of what we are looking at here, but, nonetheless, raises in my mind the importance of the work.

Those are my only comments on importance really.

Dawn, did you have others?

MEMBER DOWDING: I didn't. I actually though this was a really important issue. I guess, given the response of the developers, there is a bit of me going, wouldn't it be better to review this measure when we had all the pilot, all the data, rather than trying to make decisions on the basis of all practices?

I would be much more comfortable and more excited about looking at data variation and things if it has gone to full implementation and they are going to have a lot of data that we could look at, which would be able to provide us

with good insights into variability in practice, reliability and validity of the measure, and really understand how it could be used to measure quality of care. I think that would be much better than this effectively saying it is really interesting and important, but we need more data, when we know they are going to get more data.

MEMBER MORT: I couldn't agree more because I think, if we get to the next piece, the issues around reliability and risk adjustment, great ideas, the fact they are capturing indication and collecting a number of very important comorbid issues, but it is too soon to say. The jury is out. So, some of these things are still, I think, just as Dawn said, too early in development to really weigh-in with a definitive vote.

MEMBER BIERNER: I have a couple of comments, and I fully recognize the importance of this particular set of diagnoses that undergo surgery.

My concern is that the Oswestry may

not be the best tool. It is primarily aimed at pain, but it doesn't capture other neurological dysfunction. And you mentioned the four groups, which includes spinal stenosis.

And some patients have impairment in bowel and bladder function or weakness that may relate to either postoperative complications that occur as a result of the procedure itself or due to the disease for which they were receiving the surgery.

What is your concern about that? I don't see -- I have used the Oswestry before, and it doesn't capture some of those areas very well. What are your thoughts about that?

MS. PITZEN: This is Collette.

Although we are not putting forth these measures, additionally, we are capturing quality-of-life scores. We have been working the EQ-5D and now are transitioning to PROMIS 10. So, those kinds of measures are being captured for this patient population as well as pain scale measures or leg and back pain pre- and

postoperatively. But those aren't part of the measure that we are presenting today.

MEMBER BIERNER: Yes, I realize it is not directly, but what I am saying is, this tool, what you want to do is use it to assess the rate of -- or one of the things it will be used for is the rate of surgery varies greatly across the country and in your region as well. And I don't feel like the tool captures adequately the potential side effects of the surgery itself.

MEMBER MORT: I would think that is a very important point for the developers to consider. In fact, neurologic complications may be what push a surgeon to offer the procedure. It may also be sort of an indication, something you want to hopefully address through the surgery, but you may not, in fact, fix the neurologic complications. Or it could be a complication related to the surgery.

So, I just looked at the ODI here, and it doesn't have numbness. It doesn't have weakness. So, those would be aspects of

functional status that, to do a more complete assessment, would be good to consider including.

MEMBER THOMAS: I just have one more question, kind of a clinical question. This measure does address spinal fusion, right, spinal fusion for those with --

MS. PITZEN: That is correct.

MEMBER THOMAS: So, the real question in terms of the potential harm and the real controversial nature of this, is that with respect to individuals with disk disease or is it just in general?

MS. PITZEN: This is Collette. Let me clarify.

Patients come into the denominator of the measure by virtue of having a fusion procedure, by very specific CPT codes, and are not associated with diagnoses or other reasons. So, we are taking kind of a wide swathe of patients that are having fusions. Then, we are delineating the reason why they are having the procedure.

But the intent really was to, again, not go so narrow as the Work Group originally wanted, but to really start addressing all the patients that are undergoing this procedure.

MEMBER BIERNER: But I think that the measure that you bring forward should include, as we have said, some of these other measurements from the patient that may impact the actual success of the operation itself. And the Oswestry is somewhat limited in that way. And so it would be behoove your group, I think, to revisit that issue because I think that that is a flaw, as this is rolled out to a larger and larger group.

You had, I think, 16 orthopedic surgeons in your sample, and I think half as many neurosurgeons. But it is going to be rolled out to a much larger group of people, and you have a wider variety of skill sets or there can be greater variability in the outcomes, as you roll it out to a wider audience.

MS. PITZEN: This is Collette.

Can I respectfully disagree about the Oswestry tool in terms of some of the neurological conditions that were talked about? I mean, those kinds of conditions would prevent a patient from having full function.

If you are suggesting other tools, we would take those under consideration and share that with the Measure Development Work Group. But I guess I am not seeing the point where the Oswestry doesn't deal with function.

MEMBER BIERNER: The Oswestry is a pain questionnaire. It was geared toward assessing chronic back pain and doesn't always capture functional deficits that are more neurological in nature that could occur after this particular surgery, which has been associated in the literature with things like loss of function in bowel or bladder or -- it does have a question about sexual function.

But I am just saying that I think it could be, the measure that you are bringing forward could be improved a little by adding in

-- and I haven't looked at the PROMIS 10 -- but some other questions that might speak to those more than just the 10 or 15 questions in the Oswestry.

MS. PITZEN: Thank you. I can appreciate that. So, your concern is around the neurological symptoms that would be presenting themselves without pain?

MEMBER BIERNER: That's right.

MS. PITZEN: Okay.

CO-CHAIR PARTRIDGE: Ready to vote?

MS. ALLEN: Voting on evidence, health outcome, or PRO.

One, yes; two, no. The voting starts now.

(Voting)

All votes are in. Ninety-five percent, yes; five percent, no.

Voting on performance gap.

One, high; two, moderate; three, low; four, insufficient. The voting starts now.

(Voting)

All votes are in. Thirty-two percent, high; 42 percent, moderate; zero percent, low; 26 percent, insufficient.

Voting on high priority.

One, high; two, moderate; three, low; four insufficient. The voting starts now.

(Voting)

All votes are in. Sixty-eight percent, high; 32 percent, moderate; zero percent, low; zero percent, insufficient.

CO-CHAIR PARTRIDGE: Moving on to reliability, are there further comments? We have talked about this a little bit along the way.

MEMBER MORT: I will just make a few more comments.

I think the specifications look very clear and some of the additional comments about adding in the specific indication I think is right on target. The risk-adjustment specifications are listed, but it hasn't yet been modeled or done. So, I think there you have, along the same lines, it is a little bit too soon

to say.

And in terms of the reliability testing, the ODI has been around for a while. As a measure, apparently, it is -- I'm just looking for the comments here; it is in the notes very nicely stated. It behaves well from the perspective of reliability testing. However, the score level, that is the change in the score pre- and post-surgery, again, has not yet been subject to scrutiny. So, again, the same theme there.

There was a problem raised in the writeup about being able to get the proper denominator as well. So, that was something else that I thought was important. In other words, knowing patients who didn't necessarily complete all of the questionnaires, this is probably more on the feasibility side, but you have patients who underwent the surgery, but weren't necessarily involved in actually completing all the questionnaires.

Those were my comments.

CO-CHAIR PARTRIDGE: Sherrie Kaplan?

MEMBER DOWDING: I mean, I guess my concern is the same as the last measure, in that we don't actually have any score-level reliability testing data. We might be better waiting until they have the full dataset.

MEMBER KAPLAN: Yes, the reliability that I am looking at is that the ICCs are patient level, not practice level, because there are only four practices, correct?

And the developer hasn't done practice-level reliability testing yet, right?

MS. LARSON: This is Jasmine.

That is correct. You know, similar to the knee measure we just reviewed, we are happy to submit testing based on the information learned today.

CO-CHAIR PARTRIDGE: Okay. Are we ready to vote?

MS. ALLEN: Voting on reliability.

One, high; two, moderate; three, low; four, insufficient information. The voting starts now.

(Voting)

All votes are in. Zero percent, high; 32 percent, moderate; 21 percent, low; 47 percent, insufficient information.

MS. SAMPSEL: Okay. So, this is where we would stop. But I think, as we did similar with the measures earlier this morning, if there were any additional comments or feedback for the developers regarding what can be done differently, I think we have already had some discussion about that.

But, you know, talking further through validity, feasibility, and usability, if there is any additional guidance to the developers, they will have the opportunity to bring data back to us.

MS. LARSON: I'm sorry, this is Jasmine at Community Measurement.

May I ask just a process question?

CO-CHAIR PARTRIDGE: Yes.

MS. LARSON: It seemed to me that the group continued through voting on all of the

other criteria for the knee measure, even after failing the reliability criteria, understanding that additional information would be forthcoming.

So, I guess I was wondering why that conversation was not going to continue for this spine measure.

MS. SAMPSEL: Actually, the votes were different. With reliability, for the knee measure, the votes fell into the gray zone, which means we do continue moving forward. In this case, you have 68 percent in low or insufficient, which means the measure fails at reliability.

MS. LARSON: Okay. I understand. So, then, does that mean we will still be able to provide additional information after --

MS. SAMPSEL: Yes. Yes, we will be on the same timeline.

MS. LARSON: Okay, and everything will be evaluated at that time, assuming we progress through the criteria?

MS. SAMPSEL: Correct. If you are able to, you know, if there is additional

information for the Committee to consider, correct.

MS. LARSON: Okay. All right. Pardon my interruption. Thank you.

MEMBER MORT: I have one suggestion for the developers. If you are collecting more information about the indication for the spinal fusion, you might want to also ask whether or not non-invasive treatments were tried, such as either physical therapy or pain consults, steroid injections, just to try to get a sense for onset of symptoms, other treatments that were tried, indication for the procedure, and then you get your functional status pre and post. You will have such a wealth of important information that could add to the literature.

CO-CHAIR PARTRIDGE: Okay. Any further comments?

(No response)

If not, Collette and Jasmine, we thank you.

And I think, am I correct, that on

this measure, of all the measures we have considered so far today, this would have the highest Importance score? So, I think you have got a pretty good feeling for the sentiment on this committee that we look forward to seeing your further data.

MS. PITZEN: All right. That's wonderful. Thank you very much for your consideration and your time today.

CO-CHAIR PARTRIDGE: We are going to do one more measure on this list, which is 0631 -- oh, I'm sorry -- 2631.

MEMBER MONROE: Madam Chair?

CO-CHAIR PARTRIDGE: My vision is blurring.

MEMBER MONROE: Madam Chair, may I ask that we reverse 0688 and 2631? I have my own Board meeting at 3:30 and I have to step out, and I am a commenter --

CO-CHAIR PARTRIDGE: Oh, of course.

MEMBER MONROE: -- on 0688. So, I would ask the permission of the group to move to

that, and then come back to 2631. Is that allowed?

MS. SMITH: That is acceptable to me, if the developer is available.

MEMBER MONROE: Thank you.

I think Tracy Kline, are you on the line? I think you were --

MEMBER SALIBA: And just a reminder, I not voting on this one. So, you are one short.

MS. KLINE: Tracy Kline is here.

MS. SMITH: Great. Karen Reilly and Xing-hua Lee, are you guys there as well?

MS. REILLY: This is Karen Reilly. I'm on the phone.

MS. SMITH: Great. I want to thank you for the opportunity to present today. My name is Laura Smith. I am here with my colleague Tracy Zheng. We are from RTI International. We are here presenting as developers for this measure with our colleagues from CMS.

NQF Measure 0688 estimates the percentage of long-stay residents in a nursing

facility whose need for assistance with the late loss ADLs has increased. Increase in need for assistance is identified by comparing ratings for resident self-performance on the four late-loss ADLs, bed mobility, transfer, eating, and toilet use. We compare residents' target assessment relative to their prior assessment. This is an important measure that addresses a CMS quality strategy priority, and it is included in the CMS Five-Star Rating System.

Greater functional dependency is a risk factor for complications, such as pressure ulcer, hospitalization, reduced quality of life. Although some ADL decline may be an unavoidable consequence of an individual's clinical conditions, many risk factors may be mitigated by nursing care, multidisciplinary communication, referral for rehabilitation and nutrition services, and modification of resident's physical environment.

By monitoring and publicly reporting nursing facility performance with regard to

prevention of ADL decline, nursing facilities have the tool and incentive to focus on maintaining and improving residents' functional status.

The data for this measure is based on the minimum dataset. And for testing, we used data on all eligible long-stay residents in all Medicare-certified nursing homes nationwide, as well as previously-published studies from the development of that MBS 3.0, which was based on a sample.

Median facility-level scores for this measure were 15.4 percent and 14.3 percent in quarter two of 2014. And this measure has shown a general improving trend since quarter one of 2011.

Critical data elements for this measure show high item-level reliability and validity, with kappas above 0.95. Rasch analysis indicates that the ADL items have construct validity with items showing expected ordering with regard to the level of the difficulty to

perform each task. Items also show high internal consistency suggested by a Cronbach's alpha of 0.87.

With regard to measure-level reliability, when we look at a single quarter of data, the signal-to-noise results are low. Looking at the proportion of facilities that had scores that are significantly different from the national mean, when you look at a single quarter, we see about a third of facilities with a significantly different score. But when you look at three-quarters of data, which is consistent with how the Nursing Home Compare scores are publicly reported, we see about half of facilities have scores that are significantly different from the national mean. We also find that scores are stable from quarter to quarter when you look at that combined mean of three quarters.

We saw a low but significant correlation, suggesting convergence validity between this measure and the NQF Measure 0674,

which is falls with major injury for nursing home, and missing data do not present a threat to the validity of this measure.

Although the testing results suggest that this measure is general valid and reliable, the measure may not differentiate decline resulting from inadequate care from unavoidable decline. The measure does apply under the life exclusions, with the purpose of trying to differentiate. But there are approximately three percent of residents who died in a given quarter that were not excluded based on whether they had a prognosis of less than six months to live or were on hospice.

With regard to risk adjustment, C-statistics for the model tested were low. So, there is no risk adjustment currently applied to this measure.

There are several related measures, but none have the same focus and none target the same population. This measure's focus on functional decline is the most appropriate for

long-stay nursing home residents. And as I said before, public reporting of this measure via Nursing Home Compare provides valuable information for residents and their families.

And we thank you again for this opportunity and look forward to the discussion.

CO-CHAIR PARTRIDGE: Thank you.

David?

MEMBER CELLA: Ann asked me to lead off.

So, thank you. That was a great summary. You have touched on all the things that are relevant, and I think given it in a very nice, coherent way.

In terms of important, I mean, this is a population that is going to as a group decline. And so, the therapeutic goal makes sense, to delay decline, to avoid it where possible, and therefore have some other benefits that are likely to occur in terms of fall risk and other things that co-vary with these four late-loss ADLs.

So, I think the selection of an area and, in particular, ADLs selected are very sensible in the goal. I at first had to struggle a little bit with this goal of maintenance, but it actually makes sense, I think, with this population.

And if I understand it right -- and this may be oversimplifying -- maybe you could confirm that, basically, your sort of base rate is at about one in seven people will lose one or more functions in a three-month period. Is that --

MS. SMITH: Yes, it is at a different threshold of going up sort of two levels on at least one ADL or --

MEMBER CELLA: Okay, as defined?

MS. SMITH: Yes, as defined, yes.

MEMBER CELLA: So, the rate of decline that a facility is starting from as an average or median is one in seven, and the period of time is three months. Am I --

MS. SMITH: That's right, because the

time between the target assessment and looking back in time to the prior assessment is, on average, a quarter to three months.

MEMBER CELLA: So, if the system were a closed system, and I was trying to get maximum differentiation, I would probably want more like one out of three than one out of seven, and that would maybe make me go to six months as opposed to 90 days. I think the choice of 90 days is on the short end because you are going to have fewer events.

And maybe that decision was made because of loss in both the denominator and, then, unfortunately, the numerator if somebody dies or whatever. So, I guess would just pause for a quick answer to that. Did you consider something longer, like six months or not?

MS. SMITH: So, this measure is actually -- well, I think we noted that this was maintenance now, but it actually is based on the original MBS 2.0 version of this measure. I don't think we did consider that in this current

round because certainly we consider whether there might be provisions in this area.

MEMBER CELLA: Okay.

MS. SMITH: So, I am not sure what sort of the history of it was.

MEMBER CELLA: Thank you.

And related to that, and maybe what I projected, you know, just trying to get into your heads when you are putting something like this together, there are a lot of exclusions for this denominator, people that come in totally dependent. That, of course, makes sense. People that are near totally dependent. I guess that also makes sense. Coma also makes sense.

But, then, I start to struggle because six months' expected survival is the next reason for exclusion. And I don't know how people do that, but I know the literature on doing it is pretty bad.

So, are you just asking the person's primary MD if they think they are going to live six months, and based upon their yes-or-no

answer, they go in the denominator?

MS. SMITH: So, it does need to be based on the primary, what is in the medical record, yes.

MEMBER CELLA: But that is usually not charted, right? I mean, maybe I am wrong about that because I don't look at nursing home charts. But I don't think people usually -- I mean, they will chart if they are going to hospice. So, you will know that, and that is the next -- I can understand doing hospice and saying, okay, that is a subgroup. But, short of hospice, I guess I would remove that less than six because that could be gamed. I mean, somebody could say, "Let's put them on the less-than-six-months list," and then, they don't count. Or, if they have got 32 residents eligible, they could put three on that list, and they wouldn't have to play, because 30 is the minimum. Anyway, that is just a quibble.

MS. McMULLEN: Hi. Tara McMullen from CMS.

MEMBER CELLA: Yes.

MS. McMULLEN: It is on the medical chart, and, actually, it should follow at admission from prior assessment or prior location. We should have that type of information.

MEMBER CELLA: Okay. So nursing home admissions require charting life expectancy?

MS. McMULLEN: We take from some charts, depending on the measure. Yes.

MEMBER CELLA: That's interesting because doctors can't do that.

MS. McMULLEN: Yes.

MEMBER CELLA: I guess they're not good at it.

MS. McMULLEN: No, right. It is an exclusion that I think at CMS we have also looked at and said how reliable it is. But, for this measure and purposes --

MEMBER CELLA: Okay. All right.

And then, there was a missing value basis for exclusion which troubled me because, if

you are not good at it, then you are able to exclude people from your denominator because you are not getting the data, which seems like I wouldn't do.

What is the question?

MS. McMULLEN: It's in the administrative section, Section A. Let me pull that up.

MEMBER PARISI: It is a question on the MDS there.

MS. SMITH: It is, but the instruction is to base it on the medical record, though. So, whoever is filling it out should be basing it on the medical record as opposed to sort of it is not the MDS nurse that is making that assessment.

MEMBER CELLA: Well, it has been a while since I have published in this area, but I actually did some studies on predicting survival. Unless people have gotten better at it, going out more than a month is not good at all. I mean, the reliability, it seems like it is not worth asking.

MEMBER MORT: If I could just add to that, you said three people died in your population within 90 days?

MS. SMITH: It was 3 percent of the population.

MEMBER MORT: Three percent?

MS. SMITH: Yes.

MEMBER MORT: Even so, that is relatively low for a nursing home population. So, I think you would have some wiggle room to exclude that highly subjective determination. It may be in the chart, but, believe me, I completely agree with Dr. Cella's concerns. I don't think you are losing much by excluding that.

MEMBER CELLA: That is all I am going to say about importance/relevance. So, let's open it up for maybe Ann.

MEMBER MONROE: I just want to clarify because it looks like it says that the patients who died in that quarter were counted in the numerator. It was 31 percent. Is that a

different -- it is substantially higher than the overall incidence rate.

MS. SMITH: So, the 3 percent is the proportion of the long-stay sample that died in that quarter, but you are correct. So, that 31 percent that you are talking about is the rate, the proportion of the folks who died, the proportion of them ended in the numerator. So, there is a higher numerator-triggering rate amongst the people who died, which is not too surprising compared to the rest of the sample.

If it is all right, I just would like to address the missing data portion, which is -- while I do recognize your concern with that, what we have found is that we actually have a very low missing data rate. For this particular measure, it is only about .9 percent. So, I just wanted to put that out.

MS. McMULLEN: Yes, and I do want to address, it is in Section A of the MDS, the current MDS.

MEMBER CELLA: Thank you.

MEMBER MONROE: My only comment on importance relates also to the denominator exclusions because, at least as I see it where I observe, the line between long-term care, palliative care, and even hospice within a nursing home is getting more and more blurried. So, I wonder about those exclusions, whether they are perhaps more arbitrary. They assume more of an arbitrary nature than they really are. So, I don't know why you would exclude them in this measure.

MS. SMITH: Well, I think the intention is -- there are multiple intentions, but one is to recognize that, if people truly are at end of life, that they are going to be at much higher risk for ADL decline. And at the same time, if you do include them in the measure, are you going, especially for folks who have opted for hospice, are you going to be sort of setting up warring kind of incentives where people, the facility may not be as willing to sort of set aside some of the things that they need to do in

order to maintain function, when that may not be the person's preference.

So, I think there is just this concern about, if they are included in the measure, there may be an unintended consequence where you may not be respecting preferences at end of life.

MEMBER PARISI: So, I just wanted to clarify. You said this was not a risk-adjusted measure, but the exclusions do account for some of that, particularly in patients that have no ability to progress. Is that correct?

MS. SMITH: That's correct.

MEMBER PARISI: Okay. One more question is related to the sampling, because not every resident gets included in the sample each time there is submission, correct? Is that addressed somehow?

MS. SMITH: Actually, that is not correct, that actually there is a requirement -- well, every three months there should be a quarterly assessment done for all residents. And so, there isn't a sampling being done. And so,

we basically look at assessments that have been submitted for a particular quarter and basically identify everybody who has had -- so, there is like a slight possibility that you might not fall into it. No, you should actually end up in every -- everybody should have an assessment.

MEMBER PARISI: After a given period of time, but not every quarter? Agree?

MS. SMITH: No, it is actually every quarter, actually. Oh, but are you talking about the long-stay?

MEMBER PARISI: The MDS admission.

MS. SMITH: You're talking about the long-stay definition?

MEMBER PARISI: Yes, yes.

MS. SMITH: Okay. You're correct about it. Excuse me. This is a long-stay measure. So, you have to have accumulated 100 days in order to be included in the measure. But part of that, it really ends up being sort of a form of stratification because you want to look at a population where ADL decline is more

appropriate to be monitoring, as a particular concern for the nursing home to monitor and prevent.

MEMBER PARISI: So, one more clarification. Is there also a reflection of other outcome measures that are being collected as well, correct? Is that what I heard? Right, right.

MS. SMITH: Oh, oh, are you talking about the correlation analysis that I reference in the summary? So, what we did with that analysis, we were interested in looking at validity at the measure level. And so, one strategy is to look at whether you see correlations amongst quality measures that may have some similar underlying processes or focus for a facility. And so, we looked at how well correlated this measure was with falls with major injury.

MEMBER THOMAS: So, I will repeat that. So, I had trouble with this measure because all the other measures, pretty much, that

we looked at measured functional improvement, or many of them did. And this measure looks at the percentage of residents whose need for help or greater assistance with activities of daily living is increasing. So, it is kind of looking at patient decline in function.

And I guess there is a 20-year history of some interest groups in the disability community, in particular, who are predominantly younger disability as opposed to over age 65, who are very upset with the Medicaid program and Congress and everyone for unnecessarily warehousing people with disabilities in nursing homes, and not providing enough home/community-based services.

And they might look at this and say, well, this is a measure of poor quality. This is a measure of how, if you are a 45-year-old with MS in a nursing home, and you are declining in function, and this is measuring your decline in function, that is just proving the point that you shouldn't be there, and you need greater

engagement and services and the like. Again, I am not sure where I am going with this, if it is even a question, but how do you respond to that?

MS. McMULLEN: So, I can respond from the CMS perspective. I will try to help here. Measures like these are important to CMS because not only do we publicly report and we benchmark, and we use this type of data for care coordination and goals of care, and things like that. But we also take this data to build on different types of measures, efficiency measures, utilization measures, things of that nature.

So, in understanding an individual's complexity while they are in a specific setting, it allows us to kind of build measures that allow us to look at quality, how we can improve quality in those settings, what is going on with the individuals, look at a facility, look at their practices, improve upon those practices.

And then, now in the way of standardization, allow us to look, if they leave that nursing facility setting and they go into

some home care or they go into the home/community-based setting, what does that look like? What do those transitions look like? And as they moved, what was the change? Did they become more dependent? Did they become more independent? And how does the data paint that picture for us?

So, this could be viewed in a negative light, well, we are looking at poor quality, but, actually, all this data is used for many different reasons, from anything from payment to care plans and goals of care. So, it is not just reporting on a facility that people are -- oh, and surveyors use this as well in the QIOs. That is my boss, Mary Pratt, the Director of the Division of Chronic and Post-Acute Care.

So, it is not just used for just reporting a negative outlook. It is used for many reasons. The data is kind of recycled. And you can look at it through many different lenses, depending on what part you are at CMS, what role you play.

MEMBER THOMAS: Thank you.

MS. ZHENG: And a full-up comment on that. I think you mentioned that nowadays, given like there are more options in the community- and the home-based setting, actually, now we see this trend in terms of case mix in the population in nursing homes. Now it seems like, because of those increased options in communities, people actually entering a nursing home, and in the nursing home after 101 days will become long-stay. They are very frail, and their goal there is really to maintain function and not to have further decline, as opposed to restore function to a higher level and improve their function level.

So, as Laura said, given this population and given their risk and their goal, we think a functional decline measure is a negative measure, like higher value is that quality, but we think this is more appropriate for this population.

MEMBER LINDBERG: Thank you. Well, I

had another comment. But, first, to add to Peter's point, it seems to me that this meshes nicely with the recent ruling, basically, that Medicare beneficiaries do not have to prove improvement in function to continue certain therapies. So that, even as they may stay flat on their level of function or decline, that they should still be able to receive the therapies. So, I think in that sense this is another positive thing that meshes well with that ruling.

The question I had, though, related to the other issue around the movement, if you will, toward potentially providing curative care along with hospice and the administration's current demonstration on that. And I would just want to make sure, or I would be interested in knowing how you would look at those individuals, not excluding them, because they could have a longer period of being on hospice than the six months that is required for being prescribed hospice care. Thank you.

MEMBER BIERNER: I have one question

that I am not clear on. There are now units that are labeled as cognitive care or Alzheimer-type dementia units. Are those included in your group that you are sampling from?

MS. SMITH: Yes. Potentially, yes. There's no reason why they would not be there.

MEMBER LINDBERG: Yes. I think mine, I didn't say it maybe in the right tone, but it was a question to you. Okay.

MS. SMITH: Sorry. I think part of it was also because it is a difficult question to answer. I mean, I think it will be that case that you are talking about will be very difficult to identify when you were already sort of talking about difficulties in identifying prognosis of six months.

I think it is something that this is something that is, maybe, going to be end up being an unsatisfying answer, but I think it is something that we will just have to do some thinking about and continue to monitor. Because I think figuring out how to identify those

individuals that we are talking about, it is going to be something that is complex.

MEMBER MONROE: Another aspect of that complexity that I looked at was you kind of had a throwaway line that some people are just going to decline, but there are a lot of things that could be done. How are you distinguishing between the two, between what is kind of just natural or expected decline, and decline that results from a lack of attention to tasks that could improve the functionality?

MS. SMITH: So, the way this measure is currently operationalized, the main way that we are distinguishing is with those exclusions that we were talking about. Otherwise, the decline is being counted in the numerator.

MEMBER MONROE: Plus, you are in an exclusion. It is assumed to be fixable or improvable.

MS. SMITH: Well, I am not sure that actually that is quite how one should interpret these measures because I don't think there is

ever an expectation that the measure is going to go to zero. But higher rates are equivalent with -- tend to interpreted as worse quality. But I don't think there is any expectation. Because I think what you are talking about is basically that there would be an expectation that it should be possible to go to zero. And there is no expectation of that.

CO-CHAIR PARTRIDGE: Becky?

MEMBER BRADLEY: And if this is not the right forum to ask to respond to this, let me know. But I guess, because we had so many measures that we reviewed and so many of them were similar and somewhat overlapping, and I kind of got them all confused in my head, but there was in many of the measures that CMS is presenting the implication that they are trying to standardize the tools across settings in the post-acute setting.

But this one seems so different from some of the others, but you are continuing to want to use it and endorse it. But, I am just

curious, how does that fit into the philosophy of standardizing the measures and will this one continue to be used going forward? Because this is one of the post-acute settings that is mentioned.

MS. SMITH: So, you're right. I mean, I think that in nursing home, though, we have both the post-acute and long-term services and support populations. So, this measure is designed more with the long-term services and support population in mind. That is not entirely answering your question, and -- I don't know if Tara wants to respond to the rest of your question.

MS. McMULLEN: Yes. So, and this kind of gets into the next measure with RTI and CMS. But, yes, CMS has been, for a while, moving in the way of standardization. You saw that through the PAC PRD and the CARE tool, not the advent of the IMPACT Act. In the IMPACT Act, skilled nursing facilities are delineated among LTCHs and IRFs and home health agencies for

standardization.

This measure touches, like Laura said, upon the long-stay residents. This measure is used for the nursing home Five-Star Program. So, it is publicly reported and benchmarked. It is a Five-Star measure. So, basically, it is used to kind of report about the quality of a facility and add some sort of weight to that, to this measure, so that, basically, providers and consumers are able to make better choices about loved ones and things like that.

I digress. So, the measure was used in a different way than the actual intent of the IMPACT Act. It will continue to be used because CMS has found that it is a good measure. It is basically saying, what is going on with that person and when. At that target assessment where were they, in terms of late-loss ADLs? How did they score? And that is useful when you are comparing this measure with other types of Five-Star measures like restraint use and falls and things like that.

MS. SAMPSEL: I am mindful of Ann's obligation at 3:30, is that correct? 3:30? 3:30. And I would like to get us through this measure, if we can, before we lose her. There has been considerable mention around this table -- and I think it is a feeling that is shared by a number of us -- that the one exclusion regarding expectation of six months is disturbing. And I am going to task our technical staff, if we were comfortable supporting this measure if that exclusion were eliminated, is there any way we deal with that in our voting or do we just have to work on making it a recommendation from the Committee for the next time the measure comes to us?

MS. THEBERGE: You can conditionally recommend, and then, the developer has the option to agree to make the change, in which case they would do that and bring it back. Or the option to disagree, in which case the measure does not move forward.

CO-CHAIR PARTRIDGE: We wouldn't have

an opportunity to revote? If they came back and said, "We can't"?

MS. THEBERGE: I think you could revote if they came back and said they can't.

CO-CHAIR PARTRIDGE: Okay.

MS. THEBERGE: You would have the opportunity to revote.

CO-CHAIR PARTRIDGE: Importance.

MS. ALLEN: Voting on evidence. One, yes; two, no. The voting starts now. Ninety-four percent, yes; 6 percent, no.

Voting on performance gap. One, high; two, moderate; three, low; four, insufficient. Voting starts now. 44 percent, high; 44 percent, moderate; zero percent, low; 11 percent, insufficient.

Voting on high priority. One, high; two, moderate; three, low; four, insufficient. Voting starts now. All votes are in. 61 percent, high; 33 percent, moderate; 6 percent, low; zero percent, insufficient.

CO-CHAIR PARTRIDGE: Okay, moving onto

reliability. Ann or David, any comments?

MEMBER MONROE: I thought reliability was pretty standard. I had some comments on validity, but not on reliability. I don't know about you, David.

MEMBER CELLA: No, the same. Reliability was actually very good. The only thing I would say that there is one thing I looked at which is the stability of the facility ranks, which requires an assumption that facilities shouldn't change very much, which is probably a fair assumption. You know, looking at seeing it from time to time facilities change a lot relative to others, and they don't. Of course, if they really did, but you are assuming that they don't. And that is probably a good assuming. So, that is good. All the other reliability statistics were actually quite good. Sherrie?

MEMBER KAPLAN: Yes, I had some concerns about reliability because at the patient level it is pretty well. At the facility level,

however, the signal-to-noise analysis that was done indicates that there is a fair amount of noise in this measure at the facility level, and you can't distinguish the measurement error in the population from little perturbations at the facility.

So, while at the patient level it is good, I am troubled about the reliability at the facility level. And then, my trouble is exacerbated by the validity testing, which you found that there was a substantial stability, which is good news at one level and bad news at another. If it is not detecting fluctuations in quality of care, it is not useful for discriminating facilities one from another.

And the correlation with the other variables that were used for percentile ranks was pretty small. It is 1 percent. The R was .09. So, R-squared is 1 percent of the variation. So, that is not great news when you are using these to discriminate the care provided by different facilities.

MEMBER CELLA: Should we talk about validity now? Do you want to talk about validity or --

MEMBER THOMAS: Just a quick followup to that statement. Could you just clarify what it means that the facility characteristics -- that the measure was not particularly reliable in separating facility characteristics from noise, the population variance, what does noise mean? What are we talking about?

MS. SMITH: So, one thing, I definitely recognize your point about the signal-to-noise analysis. One thing I just wanted to point out, that I don't know whether it would actually mitigate your concern. We only had reliability signal-to-noise analysis for a single quarter of data. Recognizing that there is noise in the estimate, what is publicly reported is actually an average of three quarters.

We weren't able to do the signal-to-noise analysis for that. So, we did some analysis just looking -- we calculated confidence

intervals around the estimate for every facility. What is in the testing form is still one quarter. It is using 30 percent of the facilities are significantly different.

When we looked at that average of three-quarters, it was about 50 percent of the facilities that had 95-percent confidence intervals that were significantly different from the national mean. I know that doesn't kind of fix above .08 for the signal-to-noise.

MEMBER KAPLAN: Earlier measures developers did a nice line. You know, if it was generated from a generalized estimation disclosure of hierarchical modeling, wouldn't that be nice in this case, so that you could actually look at how good, how useful this kind of measure is, and over quarters that would make you happy in terms of there should not be random fluctuations one quarter over a whole year's worth of data, for example?

And certainly, CMS, if nobody else, has those kind of data. CMS has those kind of

data. So, that would actually give you a bit more confidence that you can use this to discriminate high functioning from low functioning. But one of the other things I noticed in your data was that 77 percent of the facilities were smack in the middle. The variation was, then, on the extremes, and that is when you plot those lines, often what you see is that these things are only merely useful to identify outliers.

So, how you are going to use this makes kind of a huge difference. You are certainly not going to use it, I don't think, at the individual facility level, but where you slice those benchmarks to do anything else, that gets at the issue of, gee, are you going to put confidence intervals around those little thresholds? How is this going to be used?

Because your signal-to-noise analyses are a little bit worrisome unless they straighten out with additional data, and your validity evidence doesn't suggest that what you are seeing

now is associated with other things you would expect to see. One percent of the variation shared in your rankings with other measures of quality around the same topic aren't really confidence-inspired.

MS. SMITH: Again, I am not sure. So, one thing about the validity data is that this is not atypical for the MDS-based measures, that they have historically not been well-correlated.

MS. McMULLEN: It should be stated that on Nursing Home Compare, like Laura said, that we are looking at multiple averages of weighted data across multiple quarters. So, it is not just assessing on one quarter for one targeted period and reporting at the patient level. It is a facility-level rate that is reported, I think, over three quarters and it is averaged.

So, you have the average for the entire country per state. So, it is a rate. You are not looking at the -- it is not a patient-level statistic. But there are some weights

applied to it, absolutely.

MEMBER KAPLAN: Okay. And the only way out of that, because you get into these tautologic loops when you are trying to figure out validity, so the only other thing you could do is maybe look at efforts to improve quality at the nursing home level. Did they do them or not? Or was there some kind of ongoing program among some or not? And then, did they move in conjunction with your expectation that this measure would reflect efforts to improve quality?

MS. McMULLEN: Yes, I think that is a point well-taken. That is absolutely what we think as well. I mean, the measure just reports on basic outcome. So, it could be used and revised and created or paired with so many other measures, so that you can look at so many outcomes. But, for what it is, it is just assessing the person at that time at that target assessment for what was going on with them. But you're absolutely right. And in the future we will expand upon this measurement set because it

is interesting to look at. I mean, it is gross indicator of so many things.

CO-CHAIR PARTRIDGE: Are we ready to vote on reliability? Or do we still have more conversation? Excuse me.

MEMBER MONROE: Did you say reliability and validity or just reliability? Because I have a comment on validity if we are --

CO-CHAIR PARTRIDGE: Reliability first. Well, reliability, let's do reliability first because I think validity may raise additional issues. Okay? Reliability.

MS. ALLEN: Voting on reliability. One, high; two, moderate; three, low; four, insufficient. Voting starts now. We are still missing a vote. Eighteen percent, high; 65 percent, moderate; 12 percent, low; six percent, insufficient.

CO-CHAIR PARTRIDGE: Moving on to validity now.

MEMBER MONROE: On to validity, one of the points that was made in the summary was about

the variation by state. And I think the Five-Star reporting is related to a national standard. And I do think state policy has significant influence on staffing, on dollars, on a number of things. And I am not sure how valid the measurement across states might be, if you have very different state policies that give these nursing homes very different tools with which to work. So, I don't how you manage that, and maybe that is not a validity question. But it certainly was an important one to me.

And the other one is related to kind of the unintended consequences of beginning to see nursing homes do adverse selection in order to only have patients where they can see an opportunity for them to improve. And I wondered how that would fit with this as well.

MS. SMITH: So, let's see, is it all right if I take the second one? I will take the second one. I will take the second first because I think it does seem like your first question is more of an implementation-type issue, yes.

And then, in terms of -- I am trying to think about whether or not this measure is useful in terms of any kind of balancing-out concerns about what you were asking. I mean, I think this measure is trying to focus on a different objective in terms of reducing, preventing ADL loss. And I am not sure that this measure is going to help with that particular concern about cream-skimming that you are talking about and adverse selection. It is still an important focus because there is going to be a segment of the population that is their goals of care aren't improvement. I mean, it does point to risk adjustments.

MS. McMULLEN: Yes, it points to risk adjustment, which we are working on those models now. In fact, we have finally found a model that actually works. At some point, we will take this back to NQF. But, I mean, from a broader sense, I think you could probably make this same argument, if you wanted to, about all the quality measures in the nursing homes set. How they

affect process, how they affect practice, and how people interpret them.

And I think that when individuals are using the Five-Stars as a means for information and knowledge, however that is, there is variation across states, and you can make the argument that states do affect that variation. But, on the Nursing Home Compare, how this measure and how other measures are portrayed are weighted-out so that there is not as much error; there is not as much sway, based on state-based policies.

So, the Five-Star itself is weighted in three different buckets. And I don't know if this is off-topic, just kind of melded together. Okay, well, it answers the first question. So, the measures aren't the only thing that weights-out those Five-Stars. You have citations, which goes back to the citation process for the surveyors as well as you have staffing, which is collected once a year and beyond that.

So, there is a complete threshold and

balancing of that. So, this measure could affect policy. It could affect practice. But, in the way that it is represented, so that we have an adequate amount of data, and the fact that it is weighted, we are hoping that it doesn't sway that so it doesn't create adverse events. But you can make that argument about every QM, I guess, that its intended purpose could be something different than what the outcome actually is truly meant to represent.

MEMBER KAPLAN: Can I ask a question procedure-wise? I just was going to ask a procedural question. So, there is all kinds of validity. There is face validity, which it is pretty clear that this has been given a lot of thought and it looks right on the surface. There is construct validity and there is discriminate validity.

For the purpose this is being put to, it strikes me that discriminate validity is where we are. You want to be able to discriminate facilities who do a good job from facilities who

don't. And I am still not hearing a lot of evidence that would support discriminate validity at this point, unless the next round of data comes in with some additional information that gives us more confidence that you are able to discriminate high-performers from low-performers with accuracy.

CO-CHAIR PARTRIDGE: David, do you have any comments on validity?

MEMBER CELLA: The best things about the validity of this, or the best thing was the correlation with falls, which was encouraging. It would have been nice to see more correlations with things like pressure ulcers or hospitalization or other things that the writeup suggests that these basic ADLs helped prevent, all of which is clinically sensible, but there really wasn't much data, although the fall data were there.

Using the Rasch Model to claim validity is kind of sketchy. It really just helps to show that people tend to lose the

ability to toilet before they lose the ability to eat. And in between, you have transfer and moving around in bed. And that they line up that way. But that is not really construct validity, at least by the way I would think of construct validity, although it was presented that way.

So, it is nice that these things are related to one another sufficiently that you can scale them in that way and consider that to be one thing, you know, like self-care. But I don't think it really gets at the kind of validity we are talking about.

The only other thing I will say -- and this gets to Len's point earlier -- is that if and when you open up the gate for more people that might die during the followup period, because you remove that requirement of not having that box checked, which I support, I was one of the ones that is supporting not having that, it does bring back the question of risk adjustment, I think, because now you are going to have a wider net of people.

You are going to have more noise. So, you don't want this unintended consequence of nursing homes killing patients that take them out of the denominator because they weren't doing well, and then, they end up looking better than they should because more people die. So, somehow death needs to be included in this metric in some way. If not, then you could have a false denominator. But other than that, nothing to add.

CO-CHAIR PARTRIDGE: My colleagues to my right tell me that, if we are concerned about the question or the issue of whether or not the exclusion should include people who are likely not to survive more than six months, this is the point in our voting where that would come up. I don't know how many of us share that. I think we may be getting a further consult.

Helen, you came in in the middle of this discussion. Is there anything you would like to add? The concern -- you missed the discussion earlier -- the concern is if we would

be very comfortable with this measure or more comfortable with this measure if that six-month exclusion were not part of the specifications. Could we, in essence, say we would vote for it, "but for" --

DR. BURSTIN: You have to vote on the measure as it is. You can certainly have the discussion and negotiation with the developers post hoc. I was just pointing out to Sarah -- my apologies, I had to give a speech in the middle of this -- but, you know, as much as all those higher levels of validity are great, it isn't actually required. You know, this would probably get a moderate, in and of itself, in terms of validity. And so, as much as we would love to get to those higher bars, I don't want to create a higher bar than actually exists for these measures.

CO-CHAIR PARTRIDGE: Are we ready to vote?

MS. ALLEN: Voting on validity. One, high; two, moderate; three, low; four,

insufficient information. Voting starts now. All votes are in. Zero percent, high; 67 percent, moderate; 28 percent, low; six percent, insufficient.

CO-CHAIR PARTRIDGE: Feasibility. Discussion? Ann? David?

MEMBER MONROE: Can you hear me? Oh, now it is red. My only thinking about feasibility was, you know -- you used the term "self-performance". I assume that means the person performs, and somebody else rates their performance. So, it is really not a patient survey. Someone else is interpreting their performance, correct?

MS. McMULLEN: Yes, that is correct.

MEMBER MONROE: And what I don't know is the standardization of that interpretation and how clear that is, so that there is the feasibility that my review of you would be -- which is not the tool, which is not the measure, but my review that creates the measure would be the same as her review of you.

And I don't know how you standardize that or how you satisfy the feasibility of people being the same, when they are all over the place in terms of training and development and expertise, when they do that evaluation. Does that make any sense?

MEMBER CELLA: I would put that with reliability and say that this is higher moderate in feasibility. I mean, it is already collected in CARE. I sure hope it is.

MEMBER MONROE: Okay. Well, sorry then.

MEMBER CELLA: But it is just whether it is collected well, is your point. I am just saying, Ann, that you could raise about all of these, and I think that is a reliability -- it is a concern and it is a reliability concern, and not a feasibility.

MEMBER PARISI: I kind of feel like I have a little of an inside scoop here, and I think everybody should be on the same page. So, the MDS and the OASIS, it is really the same

issue. The rigor that goes behind that data collection is really, it is a lot of effort that goes into educating the nursing staff, both in home health as well as in long-term care. So, I think that is an important factor.

And these data are taken from the MDS, and when we get to home care, the OASIS as well. And a lot of education and a lot of rigor goes into developing those instruments as well as implementing them. So, I think that is an important point in terms of these outcome measures. Whether or not you can take that data and are they important for improvement, some are; some aren't. And that is reflected in the discussion. But I think that is an important point that everybody needs to be comfortable with.

CO-CHAIR PARTRIDGE: Are we ready to vote? Feasibility.

MS. ALLEN: Voting on feasibility. One, high; two, moderate; three, low; four, insufficient. Voting starts now. We are missing

a vote. Sixty-seven percent, high; 33 percent, moderate; zero percent, low; zero percent, insufficient.

Voting on usability in use. One, high; two, moderate; three, low; four, insufficient information. Voting starts now. Fifty-six percent, high; 39 percent, moderate; zero percent, low; six percent, insufficient information.

Overall suitability for endorsement of Measure 0688, Percentage of Residents Who Need for Help with Activities of Daily Living Has Increased Long-Stay. One, yes; two, no. Voting starts now. Eighty-three percent, yes; 17 percent, no.

CO-CHAIR PARTRIDGE: Is it 3:25? We did it, Ann. I suggest we take a 10-minute break, if that's okay, come back at 3:35. We will pick up. We have got a block of similar measures and, then, one dissimilar measure, 2287. And it is our goal to be out of here by five o'clock.

(Whereupon, the above-entitled matter went off the record at 3:23 p.m. and resumed at 3:38 p.m.)

CO-CHAIR PARTRIDGE: So we're going to reconvene and we will start with -- let's see, what did we skip? We skipped 2631 did we?

MALE PARTICIPANT: 2631.

CO-CHAIR PARTRIDGE: Yeah, okay, which looks so much like one we've already dealt with.

MALE PARTICIPANT: Both of us thought we did it already. So anyway, editorial kind of thing.

CO-CHAIR PARTRIDGE: All right, 2631, CMS you're up.

MS. DEUTSCH: Great. So thank you for the opportunity to allow us to present on this quality measure 2631, percent of long term care hospital patients with an admission and discharge functional assessment and care plan that addresses function.

So we have a large team who have been working on this measure as well as five other

measures that you'll hear about in the next day and a half. I'll introduce just a couple of people who are going to be speaking today. But again, there is a big team behind us. So from CMS.

MS. MCMULLEN: Hi, this is Tara McMullen. I'm the cross setting lead measure developer for the Division of Chronic and Post-Acute Care.

MS. PARDASANEY: I'm Poonam Pardasaney. I'm a research public health analyst and RTI International and also physical therapist at National Hospital.

MS. DEUTSCH: And I think on the phone, Tracy Kline, are you there?

MS. KLIEN: Hi, I'm Tracy Kline. I am a psychometrician at RTI.

MS. DEUTSCH: All right, great. And Laura Smith who you previously just heard talk did a lot of the reliability testing. So my name is Ann Deutsch. I did this work as part of a contract for CMS. I'm a registered nurse by

training. I'm also certified as a rehabilitation registered nurse.

In addition to working at RTI as a senior research public health analyst, I also work at the Rehab Institute of Chicago as a clinical research scientist, and I have a faculty appointment at Northwestern University.

So first, I would like to talk -- again we've got six measures that are being proposed. And this kind of goes back to a comment that Peter made earlier today, and Becky also brought up this idea of the standardization.

So the measures that we are presenting are all built on some standardized items that are -- we call them the care function items. One of the other measure developers also has some measures related to care items.

The other thing I wanted to be sure that people were clear about is long term care hospitals. So this is a type of facility that takes care of very, very sick patients. They're referred to as chronically critically ill.

There is about 400 or so LTCHs, Long Term Care Hospitals across the country. So these are not nursing homes. They are patients, again, who have usually organ failure, a couple of organs actually failing. So they're patients on ventilators, they are very, very sick people. People who maybe 20 years ago would not have survived.

The patients then who are admitted are having conditions such as --- they have respiratory failure, cardiac failure, often have kidney failure.

So they often have functional limitations and they're at risk for having additional functional limitations that develop as part of their treatment because they are mainly immobilized, oftentimes being on ventilators or it's just very difficult for them to get out of bed.

So that's why functional assessment is really, really important in this population. In the past there was an interest in trying to help

patients recover medically. And so patients were kept in bed and on bed rest a lot, and there's been a lot of great research recently that's focused on getting people mobilized early so that their outcomes are better.

And so part of our evidence included a literature review that Poonam spent a lot of time just kind of reviewing kind of the overall outcomes that are affected by early mobilization.

So it's things like improved cognition, less delirium, improved functional status both as observed or perceived by patients, better employment, lower readmission rates, lower mortality rates, more people being able to get off ventilators, increased discharge to community homes. So that brings back something Peter mentioned earlier.

So that's kind of an overview of the long term care hospitals. So the actual quality measure in this case, it is a process measure so it's similar to what you heard about before.

We do have assessment and care plan

linked up. We -- as part of the development process, we have been actually working on this in several phases. The actual items were developed between about 2006 and 2012 they were tested.

We've had three expert panels that have focused on these measures across -- I guess it was two different CMS contracts as well as a contract funded by the Assistant Secretary for Planning and Evaluation. So we've had a lot of TEP input.

And when we first proposed looking at functional status among long term care hospital patients, our TEP who were specialists in the long term care hospital area felt that we really couldn't create an outcome measure across all the population of long term care hospital patients because the patients were so diverse.

Sometimes patients, again, are on ventilator, other patients get admitted with severe wounds. And so those patients are put onto specialized beds in order to help their wounds be healed.

And so their mobility is limited as part of their treatment and putting them on these specialized beds. So the expert panel feedback was really we're not at a stage yet in this area that we could develop an outcome measure.

You'll hear about outcome measures in our other settings. But for the long term care hospital, across all patient populations, they didn't feel that we were at that stage. That's why we felt that a process measure was the right measure to propose at this stage in time.

So the actual quality measure is that clinicians are needing to collect and then submit data on several functional items. And we include four self care items, so that's in the area of motor function, up to 11 mobility items in the motor area depending on if they use a wheelchair or walk.

We also include the confusion assessment method because delirium is a big concern among patients who are in ICU and in long term care hospitals. We have two communication

items, comprehension and expression, and also bladder function.

So the measure basically is that the assessment is conducted for these patients by clinicians on admission and discharge. And at the time of admission, we feel it's important obviously to consider whether a care plan has been put into place based on the function data.

And so how we operationalize that in this quality measure is that we're asking the clinicians to establish a discharge goal for the patient for at least one of these self care or mobility items.

We would of course hope for more, but we thought it was reasonable to expect a goal set at admission that would be the expected outcome by discharge for at least one item.

So again, the numerator is the admission data is completed, discharge data is completed, and that there's a goal. The denominator is all patients in the long term care hospital. We do not have any exclusion criteria.

The only thing that I need to qualify is that we know from our experiences in data collection in other settings that sometimes patients -- again these are very sick patients -- sometimes they basically crash.

And so if somebody is having a medical event and need to be removed from the facility, go back to acute care or go into ICU, it may not be feasible to expect the clinicians to be able to worry about whether the person's eating or not.

And so we do say that it's okay not to have discharge data if the person has an unexpected discharge. And we have very specific criteria about discharge to acute meets that criteria.

So in terms of the gap, we've done a lot of site visits as part of the Post-Acute Payment Reform Demonstration. And we found that clinicians were collecting function data here and there, but they were not necessarily always collecting the same type of information, and they

certainly weren't collecting the same information.

So it was not standardized. And as many of you know, part of the challenge with our current healthcare system is that care is very fragmented.

So you can imagine, these patients were very sick, they're going into acute care, they're going into a long term care hospital, they're also often going on to another post-acute care setting, maybe a skilled nursing facility, a rehab hospital, or going to home care.

So the idea of having standardized items that -- or function items that would be tracked across those settings is a really appealing issue.

So let's see. Tara, do you want to add anything or are you, you're good. Okay, so again, you know, we've had a lot of input. We've been working on these measures quite intensively since 2011.

And we, as part of the development

process, put these specifications out for public comment, and we got I think 22 comments from the public in the past spring.

We also put this out formally through rulemaking in the Federal Register last year. And so in April that was put out, and then we got several comments related to these measures. And this measure was finalized for the Long Term Care Hospital Quality Reporting Program in August of last year.

So I think I will stop there and open it up to questions or comments.

CO-CHAIR PARTRIDGE: And our lead discussions on this one are David, again, and Karen.

MEMBER BIERNER: Can I just ask a simple question about -- you mentioned bladder function, was bowel function or incontinence of the bowel also included?

MS. DEUTSCH: So bowel function is actually already collected as there is -- under the Quality Reporting Program there is a long-

term care data set. And that is already collected because it's a risk factor for another quality measure. So it's already actually required on admission.

CO-CHAIR PARTRIDGE: Okay.

MEMBER CELLA: Okay, well thanks Ann, for again a very clear and accurate rendering of the submission and the history. It's useful to know this is, you know, it's a really, really tough population to think about performance measures and quality indicators and including something like functional capacity.

So I, you know, I cheer the effort and I think that in that sense it makes it important. The thing that I worry about is -- kind of related to some previous discussions -- is even though no one's taken out of the denominator, with this population how many people end up incomplete, you know, because they go to the ICU or they die or, you know, they otherwise -- if you can't get that discharge functional assessment, you're excused from, right, from the

report card, from the reporting.

And I just don't know what that number is. Maybe I missed it.

MS. DEUTSCH: Okay. So let me clarify, sorry if I wasn't clear. So if somebody has an unexpected discharge, the admission data is still required and the goal is still required because obviously you don't know if somebody is going to have this unexpected discharge.

So they're only excused from reporting the discharge information. So they're included, but they just don't have to include the discharge data.

MEMBER CELLA: Oh, that's helpful. Yeah, I was going by the introductory paragraph that talked about incomplete cases, and it appeared as though they were being excluded completely. So I appreciate that clarification.

MS. DEUTSCH: And Poonam is just looking up the percent discharge unexpectedly we did, I think, report that. So we'll get back to you.

MEMBER CELLA: I guess, and I don't know what category this is in, but about the risk adjustment. They're all high risk people, right? So I guess, you know, it seems okay not to have a risk adjustment because everyone in this group of patients is a risk.

MS. DEUTSCH: Right.

MEMBER CELLA: I mean, there's like real high risk and high risk, but --

MS. DEUTSCH: Right.

(Off microphone comment)

MS. DEUTSCH: Yeah, and we do have a code that says the activity did not occur. So if somebody, yeah, so it's -- they just have to report something.

MEMBER CELLA: Anyway, we're still at the level of importance and I, you know, I'll stop talking because I think it's good that there's something in this area because it's such a tough area. So I'm high on importance.

MEMBER BIERNER: I just want to point out that more and more patients are getting

discharged to this type of facility with pressures on the acute care side to move people into other alternatives.

And so we're seeing -- this is becoming a bigger and bigger discharge disposition for a lot of acute care patients with wounds, with ventilators, and other medically complex problems.

MEMBER VAN ZYL: I had some questions about the difficulties you had looking at disparities data. I know that you mentioned specifically that you weren't able to find anything.

But the literature about disparities in long term care is pretty large. Can you tell me a little bit more about that?

MS. DEUTSCH: So you're asking, I'm sorry, about disparities?

MEMBER VAN ZYL: Yeah. I think in importance, one of the things we're looking for is evidence of disparities among different populations, and I think that there was a

specific comment that you couldn't identify any disparities among populations.

MS. DEUTSCH: So in the long term care hospitals, there's not a lot of literature about functional outcomes in particular. And so there's really no data in the literature about disparities.

And again, this is -- we're not looking at outcomes. We're just saying was a functional assessment conducted on admission and discharge, and was a care plan put together as part of the admission process.

So we actually don't have data to know whether there are disparities in terms of just doing the assessment. So we would love to know, but at this point, there's no data out there.

MEMBER VAN ZYL: Just for clarification, because this is a process measure or non-outcomes measure, is disparities data required? Sorry, my voice is a little off. Because this is a process measure and not an outcomes measure, does the NQF require

disparities data the way they do for outcomes? They don't, right?

MS. SAMPSEL: I mean, what happens with the disparities data is -- I mean, obviously we want to see it across the board as often as we can. But there are times when the data's not there. And so it pretty -- you know, there are a lot of process measures that don't have that data.

CO-CHAIR PARTRIDGE: Dawn?

MEMBER DOWDING: Yeah, can I -- I'm just a bit confused about this measure. So am I right in thinking that you've piloted it?

MS. DEUTSCH: Yes.

MEMBER DOWDING: So do you have any data to show us on the variants in the actual score across different units in long term care hospitals because I read through all of the information you've submitted and there doesn't seem to be any indication of what the score actually is, the variation in it.

So what is the percentage of patients

who have this completed? What's the range that you found in long term care hospitals? Is there -- I mean, reliability and validity is another issue, we'll maybe get to that.

But I just couldn't see anything in your documentation to actually indicate this particular measure. I saw a lot about how the CARE measures function, but I didn't see anything about whether or not a patient actually has an assessment completed and it was linked to a care plan. So where is that data?

MS. DEUTSCH: Okay, great question. So we did as part of the material report the percent of missing data for the admission and discharge. So that's in the missing data section.

And so we had 4,186 records as part of our pilot. And there were three items on admission that had missing data, that was 6.14 percent. And then at discharge, the same three items, it was slightly higher, 6.67 percent.

So I should clarify. We didn't test

the care plan part of this, that is done in other settings. So we were not able to test that part of it. But we did report the missing data.

And I should clarify that this was done as part of the Post-Acute Payment Reform Demonstration, and so the facilities were volunteers. And, you know, we would generally expect when people volunteer to be in projects that they are probably among the higher quality facilities. So I'm not sure we could generalize that that would apply to all, anyway.

MEMBER DOWDING: But the measure you're asking us to endorse includes the care plan bit.

MS. DEUTSCH: That is correct.

MEMBER DOWDING: But it's not in any of this.

MS. DEUTSCH: We don't have data on that. That is correct at this point.

CO-CHAIR STILLE: I just had a comment about -- we had a fairly extensive discussion about care plan this morning and how reliable

data were in detecting a care plan.

One thing that's actually sort of nice about this is that there's a measure that links having something in the care plan that's related to the assessment, which wasn't in this morning.

But I think, you know, data about how possible is it to measure both of those is going to be really important to look at value.

MEMBER MORT: I apologize if you already mentioned this, but what are the actual data elements or tools you'll use to assess function, self care, mobility, cognition, communication, and bladder. But is there a standardized CMS tool kit that you're implying, or is there a choice? I apologize if you mentioned that.

MS. MCMULLEN: Yes, so the items themselves are derived from the CARE tool, which came from testing that occurred in the PAC PRD, the Post-Acute Care Payment Reform Demonstration.

So the Post-Acute Care Payment Reform Demonstration derived out of the Deficit

Reduction Act of 2005. So basically the Deficit Reduction Act mandated that this demonstration occur to see if anything like standardization in post-acute care settings and acute care settings was possible to be able to assess patient complexities and, you know, to look at payment and things like that.

So from that payment reform demonstration came the CARE tool. And from that CARE tool we have many sections, domains. But one domain or one section is the function section.

So we used items from the function subset of the CARE tool, and that's how we developed these measures.

MEMBER MORT: And the items are listed, I think, in Table 1. But is the assumption that all of those will be assessed? I just don't know how the CARE tool works.

MS. MCMULLEN: Yeah. So there's four self care items: eating, oral hygiene, toileting hygiene and wash up your body. So we did not

include dressing items.

And that was based on our testing that in long term care hospitals, those items -- you know, patients were very sick, they were wearing gowns and so it's really not fair to assess whether somebody can put on shoes or not. It's just too hard for them.

So we only included the items that we thought made sense for that population. In the area of mobility, we have quite a few bed mobility items.

So we have roll left to right, sit to lying, lying to sitting on side of bed, sit to stand, chair to bed transfer, toilet transfer. There is different walking distances. If somebody's walking, they can otherwise skip, there's a couple of wheelchair distances, otherwise they can skip if they don't use a wheelchair.

We have the confusion assessment method which is a published instrument. We have a comprehension or understanding item, an

expression item, and then the bladder continence item.

I do want to highlight that there's a six level scale for the CARE tool, six being that the person is independent, level one being the person is dependent. If an activity does not occur, for example somebody does not walk at this point in time, they just record the reason that the person wasn't able to do it.

Maybe the person's too sick, there's a medical reason, or maybe the person refused. So there's special codes for somebody -- the clinician to document that the activity was not attempted, and again the rationale.

So if they put the code to say, you know, this wasn't attempted because it wasn't safe for the person to get out of bed, they get credit for that. All we're asking is that there's a response for each so we know that they considered doing that assessment item.

MEMBER MORT: Thank you.

CO-CHAIR PARTRIDGE: Peter?

MEMBER THOMAS: Again, what comes to mind just from a layperson's perspective I guess is, you know, what LTCH wouldn't be doing this? And I guess that gets to patients who come to them for treatment. And so I guess it goes to the issue of whether there's enough variants to make this measure really worthwhile, you know, conducting.

Are there really LTCHs out there that take patients and then don't have any assessment and any plan of care that they've got to treat the patient?

MS. DEUTSCH: So, great question. When we did the Post-Acute Payment Reform Demonstration, we saw variability in the types of items that were assessed. But certainly, you're right, a lot of patients were seen by therapists.

I think, you know, we don't know enough to know that that's happening. And so this measure could help document that. But whether it taps out, you know, soon and we should really move to outcomes more, that's I think a

great question.

CO-CHAIR PARTRIDGE: In order, David, Liz, Becky, Sherrie. Oh, Liz went down.

MEMBER CELLA: There we go. So it's really just to now clarify a couple of things because you actually mentioned it, Ann and then Chris, I think you restated this and it was not my impression that the assessment -- that the actual functional care had to be linked to the assessment, that the goal could be just any goal and it did not have to be linked.

I didn't see that anywhere in the document. Is it true that there needs to be a link as we had in the discussion this morning?

MS. DEUTSCH: Yeah, so basically let's say, I don't know, roll left to right, somebody might be admitted at a score of level one and the goal is that they would improve. So you would link up, you know, that the goal is this item and the goal is for them to get, I don't know, level two, they would improve in independence.

So I think that's what you're

referring to, that the function item is here and then there's a goal tied exactly to that item. So that's true.

MEMBER CELLA: Okay, that sounds great. But the performance measure's not written that way. I mean, it's written that you have to have both, but not that they have to be linked. Whereas this morning we saw a performance measure that was actually written that way.

MS. DEUTSCH: Well, so maybe it wasn't clearly worded, sorry. But what we intended was that you score each of the items that I listed out. So for example, you would score roll left to right, and then you have one or more goals for each of the self care or mobility items. At least one of those items that you score.

MEMBER CELLA: Okay. Well, in my mind that actually elevates the importance as opposed to be --

MS. DEUTSCH: Oh, okay, well good.

MEMBER CELLA: But maybe I want to encourage you to rewrite the measure so that it

actually shows the link in the terminology of the measure, and the way it's described because it's actually when you read into, like, actually how you get the number, there's nothing in there that says there has to be a demonstration that there's a connection.

CO-CHAIR STILLE: I was assuming just in the first sentence it says brief description, you know, a care plan that addresses function.

MEMBER CELLA: Yeah, but then -- well okay, anyway.

CO-CHAIR STILLE: I don't know. I don't know.

MEMBER CELLA: I'm glad to hear that there is a link because that strengthens it. The other thing is you mentioned the six percent missing data. But what's the percentage of documented non-adherence -- documented non-assessment because if somebody documents that it wasn't assessed, that actually counts, they get credit for that. So how often does that happen?

MS. DEUTSCH: You're asking what

percent of time the activity did not occur?

MEMBER CELLA: So if I'm reading it right, if the provider says, you know, I couldn't assess it, patient refused, patient was sleeping, whatever, it didn't work out, they get credit for it because they documented that they didn't get it.

MS. DEUTSCH: Right. So sleeping is actually not a reasonable rationale. But you know, somebody refusing is reasonable and they would get credit that they tried to assess the patient.

MEMBER CELLA: How often does that happen?

MS. DEUTSCH: So it really varies by item. I can certainly send that after this meeting. I actually have a report I can send you graphics by setting, how often the activity did not occur.

But I can tell you, like, things like stairs, which is not on this measure, and dressing, were very commonly not assessed, and

that's why we didn't include them as items.

MEMBER BIERNER: Because, as clarification, what is the length of time and the look back that this is being -- is this assessed the last 24 hours, in this week, like, the best performance in the last 72 hours? I have the same issue with the FIM that I have this issue too.

MS. DEUTSCH: Great question. So the instructions were if the patient was admitted before 12:00 noon, there was a two day assessment period. If the person was admitted after 12:00 noon, they had a three day assessment period on admission.

Most people were admitted after 12:00 noon, so it's generally a three day assessment period. And it's mutual performance. At discharge it's any time during the last three days. Mostly it's during the last day or so. But we did give people the option of three days given weekends, therapists aren't necessarily working every day.

CO-CHAIR PARTRIDGE: Becky?

MEMBER BRADLEY: Thank you. I guess I'm having a little trouble figuring out how this would be a quality measure. I know it's a process measure. But it's -- the way I understand it it's did they do an assessment, did they do a plan.

But they could pick -- a clinician could have two patients that are pretty much the same and pick different items to focus on as the goal. And so I'm confused as to how that rolls up to some type of benchmarking or comparison to make it a quality measure.

MS. DEUTSCH: So our technical --- so just to kind of recap Becky's question, so it's basically how is this a quality measure just doing the assessment and the care plan.

So our technical expert panel, which is our LTCH experts basically said that given the heterogeneity of the types of patients, it was really hard to think about trying to do an outcome measure or hold people accountable to goals.

I mean, that was something that I think we would love to be able to do potentially in the future. And actually tomorrow I will be presenting an outcome measure that's focused on patients who are on ventilators.

So it's a very specific sub-group, but in terms of having a measure -- a quality measure that would work across the entire LTCH population, I think we just don't know about the diversity of patients.

Also, there's major payment reforms that are happening in the long term care hospitals. They are going to be paid differently in the future. And so probably the types of patients admitted will really change a lot.

And so I would worry that if we did create an outcome measure, at this point in time, that applied to all patients now, it wouldn't necessarily work well in, I don't know, five, ten years whenever that shift happens. Does that help?

MEMBER BRADLEY: It helps.

MEMBER KAPLAN: I want to follow on Dawn's point about you've done a pilot study at the facility level but we don't know what the answer to that is in terms of means and variability, in terms of a performance gap.

Help us understand why that wasn't done because things like -- my concerns would be the same as the earlier one. You know, you guys have -- if nobody else has data on this, you have data on this to help us understand what the facility level on reliability is in terms of intraclass correlation coefficients, what the other kind of measures do you have available to help us understand why we don't see any of that.

MS. DEUTSCH: Okay, great question. So first of all, unlike the other post-acute care settings, long term care hospitals only recently implemented a clinical assessment data set.

That was as part of the Long Term Care Hospital Quality Reporting Program that started in October of 2012. So it's only recently that actually, pressure ulcer data and some function

data were available other than claims.

And so obviously this is a very clinical issue. And so there really aren't a lot of data out there at this point. So our pilot study, we did collect the CARE data on admission/discharge. We did not collect the CARE plan data. That was something that was made a decision after that got started.

MEMBER KAPLAN: So there are data on the front end of this measure at the facility level. So how are we to evaluate a performance gap with no data at the facility level?

Are we -- because all of the information is at the patient level. So I'm a little bit nervous about something that's going to be used. The attribution is to the facility with no evidence that there's a performance gap at this facility level.

MS. DEUTSCH: Right. So we did actually submit some supplemental information that talked about our experiences as part of the Post-Acute Payment Reform Demonstration in that

there was variability in terms of the items that were being collected.

And part of what we also feel is important is that there would be this standardized assessment data collected across settings potentially. And that would be helpful in terms of care coordination.

So there would be a common language of function just like if I told you somebody's blood pressure, you would automatically know, you know, whether somebody was in an IRF or a SNF or long term care hospital.

So standardization of assessment items is kind of part of what we're interested in also as part of this. So that should improve care. I'm sorry, I should --

CO-CHAIR PARTRIDGE: Dawn?

MEMBER DOWDING: Okay, I'm still going back to my same point. I'm really concerned that you're asking us to endorse a measure where you actually don't have data at all on part of it.

Like, you've said twice you haven't

collected the data on the care plan bit of this measure. So we can't -- I mean, I'm sort of sitting here thinking it's quarter past four. We're discussing a measure that we actually can't endorse because you haven't collected the data on part of it.

And I'm really, I don't know, maybe I'm just -- it's quarter past four and we've had to look at a lot of them. But it's like how can we evaluate reliability, validity, do anything with this measure when a key part of it, the key that makes it important is the link between assessment data and the care plan and you haven't got the data on the care plan.

I just -- I'm really sorry but I just think we have to say is it worth continuing with this discussion because we can't endorse it. We haven't got the data to endorse it because it hasn't been collected yet. Or am I missing something? Am I just totally on the wrong page?

MS. MCMULLEN: So from the CMS perspective, I think it's a point well taken. As

a researcher and academic, my other hat, I understand.

But in the development of process measures, we use these measures to be able to collect the data to make outcome measures, to make these measures where we are able to publically report and benchmark to a degree.

So the items would be nested within the LTCH long term CARE data set. That's the data set that Ann was just talking about. And we use these simply to collect data at this point because the data is not available. I get your point, but --

MEMBER DOWDING: But that's fine, but the NQF is not -- you're going to collect the data anyway. We're being asked to endorse a performance measure.

MS. MCMULLEN: Right, so --

MEMBER DOWDING: And we don't have the data in which to do that. So yes, collect the data and come back with the data so we can endorse it. But we can't -- I mean, unless I'm

misunderstanding the role of the committee.

CO-CHAIR PARTRIDGE: I think it's very close to getting ready to vote on importance. But I want to see if Peter or Brian were addressing importance -- issues related to importance before we take that vote. And Karen, I'm sorry.

MEMBER LINDBERG: Okay. Yeah, I just wanted to clarify. Have you had discussions about how the measure could be used for cost containment or fraud detection?

MS. MCMULLEN: Yeah, so in the Post-Acute Care Payment Reform Demonstration, data was collected for more than just looking at quality, but really looking at efficiency and utilization being prediction type models and things like that.

At this point with this data, we have not had that type of conversation. Of course, you have the IMPACT Act which mandates that we develop a resource and measures like Medicare spending per bene.

And you will have a function measure. So at some point those worlds will collide where you're looking at attributable episodes and what's going on with that individual within that episode.

But at that point, we have not discussed that. But that's definitely the direction we're moving in.

MEMBER BEVANS: I want to support Dawn's statement and clarify, I think for myself and maybe some other members of the committee that we are not being asked to endorse the care measure.

We are being asked to endorse a process measure that is about the administration of the care measure as well as the development of a care plan process, not the psychometric properties and importance of all of that of the care measure itself.

Not to say that that isn't obviously an essential component of this. It is not a sufficient component of this measure. Being a

process measure, we don't have sufficient information to be able to, you know, make an informed decision about this measure at this time. We're missing half of the information.

CO-CHAIR PARTRIDGE: Okay, on to vote. I'm sorry.

MS. MCMULLEN: -- about time limited endorsements and if those things were feasible based on data collection and coming back to the table with further data collection for care plans.

MS. THEBERGE: We're not doing time limited anymore.

MS. MCMULLEN: Okay. Thank you.

CO-CHAIR PARTRIDGE: Okay. Nadine?

MS. ALLEN: Voting on evidence, one high, two moderate, three low, four insufficient evidence, five insufficient evidence with exception. Voting starts now. All votes are in, six percent high, 28 percent moderate, zero percent low, 50 percent insufficient evidence, 17 percent insufficient evidence with exception.

(Off microphone comment)

CO-CHAIR PARTRIDGE: Okay. B.

MS. ALLEN: Performance gap, one high, two moderate, three low, four insufficient. Voting starts now. Six percent high, 11 percent moderate, 17 percent low, 67 percent insufficient.

MS. SAMPSEL: So at this point, I mean, the measure fails, we don't move forward. But as we did with the earlier developers -- you know, and I know we've had some discussion.

But I think what I've heard from the committee is, you know, what you want to see moving forward for additional data which could be submitted before the end of public comment for reconsideration and re-vote would be those -- you know, the additional data that you have as well as more information regarding the care part of the process measure because we want to tie those together.

Were there any other comments, considerations that the committee would like to

ask CMS and RTI to provide?

CO-CHAIR PARTRIDGE: If not, we're going to thank you very much. And let's see where we are. We are on -- aside from the fact that we're brain dead.

(Off microphone comment)

(Laughter)

CO-CHAIR PARTRIDGE: 0701 is the one remaining. You're up.

MS. SAMPSEL: Kate, are you still on the phone?

MR. LICHTMAN: Steve Lichtman still on the phone. I'm one of the developers.

MS. SAMPSEL: Okay. So we are ready to move on to 0701 and the developer is the American Association of Cardiovascular and Pulmonary Rehabilitation. So if you could do as brief as possible measure introduction.

MR. LICHTMAN: Yes.

MS. SAMPSEL: You know, at the same time make sure you hit your key points.

MR. LICHTMAN: Yeah. I fully

understand. We've been on the phone listening to you guys since about noon.

MS. SAMPSEL: Sorry.

MR. LICHTMAN: So it's been a fascinating listen, let me tell you. Hi, I'm Steve Lichtman, I'm a past president of AACVPR, that's the American Association of Cardiovascular and Pulmonary Rehab.

And I'm the current lead on the pulmonary rehab performance measure task force. And we're presenting to you today the functional capacity in COPD patients before and after pulmonary rehabilitation.

And I want to thank NQF and the Committee for considering our submission and letting us onto this conference call to speak about it.

Kate Murphy is also on the call, she's our staff person in charge of our task force from AACVPR. Dr. Marjorie King is also on the call, she's also a past president of AACVPR and she's the current chair of the quality care committee

under which the performance measure task force falls.

And Dr. King and I both work at Helen Hayes Hospital in New York, and we've been running the pulmonary rehab program for over 20 years now. She's the medical director and I'm the program director.

And finally on the call is Gerene Bauldoff who's a professor of clinical nursing at Ohio State University College of Nursing. And she's been working in pulmonary rehab for over 18 years. And she's also on our task force.

A little bit about pulmonary rehab. And I'll try to make this as brief as possible. Pulmonary rehab is a low cost, highly efficient, evidence based program that's been shown to improve function, quality of life, decreased dyspnea, decreased COPD exacerbation, and decreased rehospitalizations.

It's typically run in a group setting, the cornerstone is physical conditioning with many, many different devices used. Supplemental

oxygen is used, oximetry is used to make sure patients don't desaturate.

There's also a strong educational component to the program, and many programs also use breathing retraining methodology. It's recommended by the Gold guidelines for moderate and severe COPD patients as standard practice for treatment.

And a cornerstone of pulmonary rehab that's really related to all the changes is the improvement in functional capacity. This really relates to all the other outcomes that I've talked about in pulmonary rehab.

This measure that you guys are hearing today was endorsed by NQF in 2011. And it represents a clinically important measurable outcome that's vital for pulmonary rehab programs to utilize.

And the measure is defined as the percentage of patients with COPD who increase by at least 25 meters as measured by the six minute walk test.

And the six minute walk test is a valid and reliable standard test that's used in pulmonary rehab to assess the functional capacity and the functional changes of our patients. And we chose 25 meters because that is the minimal important difference that's been identified in the literature over and over.

The measure was tested utilizing two different methodologies. One we used a group of pulmonary experts, 32 experts from around the country and internationally replied to a questionnaire that looked at the reliability, the ability to differentiate quality programs, face validity, how we defined our numerator and denominator, are there any negative consequences to the measure and looking at our exceptions.

Overall, they used a Likert-like scale. And overall on all the questions, they were well above four out of five on all the categories, demonstrating an excellent response from our experts.

Then we also tested the measure using

the only nationally available database looking at outcomes in pulmonary rehab, and that's the AACVPR National Data Registry.

We used a one year period from August of 2013 to August of 2014. And in the definition of the measure, you look at change in the six minute walk test, pre and post pulmonary rehab participation, which is generally over a three month period with a minimum of ten sessions attended by the patients.

And when we examined this data, we found -- and we also looked at the raw data and we also looked at the data used in the Charleston Comorbidity Index to restratify the models also.

And there was no significant difference in the outcomes whether we used these Comorbidity Index as a covariant or whether we looked at the raw data. So there seems to be no impact of other comorbidities on our outcomes.

And we utilized over 2,668 patient records, pre and post pulmonary rehab. This represented 121 programs geographically

distributed across the country. To make it short, you have the demographic data, you have the distribution data in your report so you can look at that.

And to make a long story short, what we found is there were very few exceptions where patients didn't have pre or post data. And we did find a gap, and the gap was that 21 percent of the patients don't meet the minimal important difference in the programs across the country.

So there's a lot more analysis in the data, and I'm sure you'll question us about that, so I'll keep this short.

And in conclusion, what we really are looking at, what we really think this performance measure will allow programs to do, A is to guide them in what is important to measure in a pulmonary rehab program.

AACVPR also runs a program certification process. And we have found in that program certification process that there are many programs across the country that are fairly

ignorant in the use of outcomes, how to use them, and what to measure.

So this will initially serve as a guide for one of the most important outcomes to measure in pulmonary rehab. It will allow programs to establish or allow us to establish program quality by looking at the change in scores of a valid and reliable, clinically meaningful assessment of functional capacity.

AACVPR in the near future will be releasing benchmark data on six minute walks in pulmonary rehab, it will allow programs to compare themselves to benchmark, and most importantly will allow programs to develop quality improvement plans if they're not meeting the 25 meter change in a large number of their patients as compared to national benchmarks.

To make a long story short, that's what we've done with this measure.

CO-CHAIR PARTRIDGE: Thank you. Sherry, do you want to lead off our discussion on importance?

MEMBER KAPLAN: Yeah. Help me understand who the target of inference is, whose quality of care are we measuring? The data you provided are, some of the data to support the quotes validity are all patient levels.

There's no quality programs or physician or facility level information in at least what I could find. So help me, who is this supposed to be used to evaluate, what's the performance assessment?

MR. LICHTMAN: The evaluation of each individual patient is on a patient level. However, what we intend this performance measure to be utilized for is for programs to be able to evaluate their own quality once they understand the process and the importance of measuring the six minute walk data performance.

MEMBER KAPLAN: Okay. So now my question is what's a program?

MR. LICHTMAN: Okay, pulmonary rehab programs are outpatient programs, they're run across the country. We estimate there's

probably, oh, 900 to 1,000 pulmonary rehab programs across the country.

They're delineated by a process of who can participate in the program. Medicare currently allows moderate to severe COPD patients to participate in their programs, private insurances have a little more open guidelines, and some states allow other types of patients to participate.

But across the country it's standard that patients with COPD are reimbursed for participating in pulmonary rehab. It's generally a 12 week program run two or three times a week anywhere from an hour to two hours of rehabilitation.

There's exercise training just like you would see in cardiac rehab or a gym. The only difference is we are monitoring SBL2 and a lot of the patients are on supplemental oxygen.

There is a breathing retraining class, and then there's an education component. And those are all required by Medicare to be standard

components of the pulmonary rehab programs.

MEMBER KAPLAN: Okay, so --

MR. LICHTMAN: Typically run, oh go ahead. I'm sorry.

MEMBER KAPLAN: Oh, that's all right. Sorry, finish.

MR. LICHTMAN: It's typically run by some combination of respiratory therapy, exercise physiology, nursing, physical therapy, and sometimes occupational therapy. Multi-disciplinary in nature, and it's designed to improve the function and the quality of life of the patients enrolled.

MEMBER KAPLAN: Okay, so right now though we don't have any evidence of between program differences or between program reliability, variations that would be attributable to, for example, how precise, how reliable or reproducible these scores are at the facility or program level.

MR. LICHTMAN: That is correct. The only national database, the AACVPR database when

we submitted this, even though it had over 2,000 patient records, when you spread them across the programs, we really didn't yield enough data with each program to look at that.

However, the six minute walk test and functional capacity -- and Gerene can speak to this a little more -- has been shown to be extremely valid, reliable, and important in the established literature. And we submitted that in the evidence.

MEMBER KAPLAN: Okay. So reliable and reproducible scores, what you've actually provided us in terms of the agreement of expert panel is what we call content validity. That is did you sample correctly from the domain of observables.

And that really is more or less face validity, are they right -- it's not a reproducible score at the facility level. So right now, we don't have any information at the facility level on which to endorse this.

But that's the intent for us because

we can't -- without that information, I'm concerned about what actually we're being asked to endorse. Furthermore, without some evidence that this is hooked up at the facility or program level with something else, for example, for construct validity that you would think would actually distinguish facilities or programs in terms of variation quality, you don't have that either.

So I'm a little bit lost about exactly what it is we're being asked to endorse.

MR. LICHTMAN: Well, we did ask the expert panel if this would differentiate between quality programs, and they strongly agreed to that.

As far as what we're being asked to endorse is the program is to have a performance measure to follow so that they can begin testing and measuring this in the appropriate fashion.

MEMBER KAPLAN: So this is a question, I guess, for the NQF staff.

CO-CHAIR PARTRIDGE: Yes. I think, do

we want to have a two minute pause and let you confer?

MS. SAMPSEL: No, I mean, so your points are valid. However, the way that the criteria are written, this is not a patient reported outcome measure.

So on an outcome measure, all that is required is either the item or the kind of patient level result, or the measure level results.

MR. LICHTMAN: That's what we were --

(Simultaneous speaking)

MS. SAMPSEL: Or, you know, that type of testing. So frankly, they did provide the amount of testing information required for an outcome measure. But if the patient --

MEMBER KAPLAN: So no, because --

MS. SAMPSEL: At the item level, so it could be at the scale level of what they were -- or the tool level of what they've done, they can provide that for this measure. And that meets NQF criteria.

MEMBER KAPLAN: Okay, so this sounds like one, if we had a tiered system where this would be like the Phase 1 and the FDA approval level, this is to go out and start collecting data rather than use it to discriminate programs because right now it doesn't sound like we have enough evidence that it's valid or reliable for that purpose.

And there is no such thing as a valid measure. They're valid for populations and purposes only. So if we use that as a criteria, I'm still a little bit flummoxed about how much evidence we have to support the reliability and validity for the purpose that it's intended to be used for unless it's just at the patient level.

MR. LICHTMAN: Well, the measurement is at the patient level and it's for programs to measure their changes in functional capacity. And then in the NQF application it said what is your plan for reporting this in the future.

And part of the plan would be to, in the future, have programs compare to benchmarks

that will be available.

MEMBER KAPLAN: See for me, Sarah, this is one of those tiered approval problems that we don't have in place yet.

MS. SAMPSEL: Well, and I think the best mechanism, you know, kind of as a committee to work through some of this is going to the algorithm and going through the step by step process which you know kind of takes you through the concept of in the first question when thinking about validity is are the measure's specifications consistent.

And then it goes down to was empirical validity testing conducted using the measure as specified for the applicable tests, and then you do the yes or the no. And if no, you go to face validity, et cetera.

But I think what you're getting to, Sherrie, is then you go down to the fact that was validity testing conducted with computed measure -- performance measure scores for each measured entity, the answer's no.

So with the NQF criteria, the next question is was validity to that testing conducted with patient level data elements. And if yes, you have options to rate as moderate to low.

So, I mean, it does kind of go through that step process. I think this is an area where in some cases with these types of measures we're in a little bit of untested grounds. And we'll look for your feedback on it, but we still need to kind of go through the process.

CO-CHAIR PARTRIDGE: Further discussion?

(Off microphone comment)

CO-CHAIR PARTRIDGE: On importance, yes. I have to say one thing. I did like the fact that various threshold as opposed to, we aren't quite sure what the right level is. Here we've got a threshold. Peter?

MEMBER LINDBERG: Again, I'm not a clinical person. But it strikes me as odd that the best way to measure pulmonary function is

through a distance, a walking test based on time that it takes to cross a certain distance. I'm just surprised by that. Is that --

MEMBER BIERNER: It's actually not surprising. It's a measure of function. You can improve someone's function and their physiological parameters may not improve that much. But their function can improve because they're debilitated from their disease.

And so this is true in cardiac as well as pulmonary rehab. You may improve functional measures like walking ability and endurance, it may not change other physiological parameters on it.

MEMBER LINDBERG: I don't know the percentage here, but what about persons that don't walk well, can't walk, have other ambulatory issues that they're dealing with?

MR. LICHTMAN: The six minute walk test is valid for patients who use assisted devices. So we can include them. And there are exclusions where patients who can't ambulate or

would be a risk would be excluded.

However, quite honestly, when patients come to outpatient pulmonary rehab, they are almost always ambulatory. Outpatient pulmonary rehab, because it's an outpatient program, the patients really need to have a minimum level of function in order to benefit from what we do.

And virtually all the patients can do a six minute walk test. They might not do the six minutes. They may have a very low value. You don't have to complete the six minutes for it to be a valid test.

And it could be for various reasons. But as long as they can attempt the walk test, it's considered a valid outcome.

MS. KING: This is Dr. Marjorie King. The six minute walk test is used in research to assess differences in outcomes, including in patients with pulmonary rehabilitation.

It was used in the National Emphysema Treatment trial which basically showed that pulmonary rehabilitation is better than some of

the surgical techniques used to treat patients with moderate to severe COPD.

So it is a tool that is used within pulmonary rehabilitation and has been valid and reliable in this population. It's also used individually clinically to assess improvements in patients in both inpatient and outpatient settings for heart failure or for COPD rehab.

MR. LICHTMAN: And the literature has been using the six minute walk test to assess function in COPD patients probably since the mid '80s. And there was a huge Medicare sponsored trial called the National Emphysema Treatment Trial that was done right around 2000.

And that was the only randomized, large scale examination of different treatments for COPD. Six minute walk test was a primary outcome to that study.

CO-CHAIR PARTRIDGE: Okay, importance to measure and report. Nadine?

MS. ALLEN: We're voting on evidence, one yes, two no, voting starts now. We're

missing one vote.

CO-CHAIR PARTRIDGE: Yes, Brian left.

MEMBER MONROE: I heard none of the discussion. So I won't vote.

CO-CHAIR PARTRIDGE: Oh, you didn't vote.

MS. ALLEN: Oh, you didn't vote. So we are just --

(Off microphone comment)

MS. ALLEN: Eighty-eight percent yes, twelve percent no. Voting on performance gap. One high, two moderate, three low, four insufficient. Voting starts now. Twenty-four percent high, forty-seven percent moderate, zero percent low, twenty-nine percent insufficient.

Voting on high priority. One high, two moderate, three low, four insufficient. Voting starts now. Forty-one percent high, fifty-nine percent moderate, zero percent low, zero percent insufficient.

CO-CHAIR PARTRIDGE: Okay, reliability next. And Sharon?

MEMBER CROSS: Just a quick question in looking at the measure worksheet before we go any further. I just want a clarification for myself.

This is an endorsement maintenance, meaning that this had already been moved forward in the past, correct? So is there something that we normally would see or that we would know as to what's changed or if there has been any changes since it was last endorsed, or is that not something that's important for our committee?

MR. LICHTMAN: I can tell you that, if it's okay.

CO-CHAIR PARTRIDGE: Go ahead.

MR. LICHTMAN: Okay. Basically, the entire testing form is new. The survey of the experts is new, the statistical evaluation from the pulmonary rehab database from AACVPR is brand new, and the literature has been updated.

CO-CHAIR PARTRIDGE: Further discussions on reliability? Sherry? And then David.

MEMBER KAPLAN: I'm beginning to sound like an old wheeze on this issue. But you know, again, without some further guidance about exactly what it is we're approving here, I get it that this is a very important test for, especially for people with cognitive deficit who can't answer questionnaires, you know, can they function, can they walk, and usefulness at the clinical level for improving, you know, for taking care of an individual is very supported by the evidence provided.

Its use for a quality measure, however, at any level other than the patient level which I can't imagine how you would use it at the patient level, is problematic for me because again, I'm not seeing evidence.

And if this has been around for a while, is there any evidence that, does the change alone mean that now we're approving a change for continuation for the prior uses or we being asked to endorse something that as the developer said, going to use it to kind of

compare facilities and gather data for reliability and validity tests?

MS. SAMPSEL: So I think this still goes back to the same issue. I mean, it's a valid question especially over time when this has been a measure in use for some time.

But when you go back to the NQF criteria of this being an outcome measure, we are just looking, you know, when you go through the criteria your choices are going to be moderate or low based on the fact that they didn't provide the measure level, reliability, more validity testing.

So that gives you the option below of then deciding from what they did provide at the item or patient level, is that sufficient for you to make a low or moderate decision.

MEMBER KAPLAN: For the purpose of?

MS. SAMPSEL: For the purpose of moving it forward for endorsement.

DR. BURSTIN: Again, ideally we'd love to see it at both levels, it's not required. But

it is a measure that, it's one of the few measures in use in the rehabs, in the sort of cardiac/pulmonary rehab space.

MR. LICHTMAN: It's almost the only measure in use in pulmonary rehab, correct. Cardiac has others.

MS. KING: This is Marjorie King. Oh, I'm sorry. I just wanted to mention that a six minute walk test is similar to checking a blood pressure for someone who does it.

There are specific, standardized ways that you do it, that you follow, criteria that you follow in order to perform the six minute walk. It's a very standardized tool, measurement tool.

MR. LICHTMAN: Yes, there's an entire American Thoracic Society guideline that outlines precisely how to do this. And I think that's one of the reasons why in the evidence form with all the previous literature showed to be extremely valid and reliable because it's standardized.

It's not just go walk down the hall.

There are specific areas, there are specific measurements and there's even a script that the clinician should follow.

Additionally, AACVPR has a toolkit up on their website that clinicians can access that goes through where to find the instructions, how to do the test, what's the minimal important difference, et cetera.

So that's all available. We put that in the appendix. That's all available to the clinician.

CO-CHAIR PARTRIDGE: All right. David, did you have any further questions? No? Okay. Then I think we're ready to, we are going to vote ready or not on reliability and validity.

MS. ALLEN: Voting on reliability, one high, two moderate, three low, four insufficient. Voting starts now. We're still missing some votes.

(Off microphone comments)

MS. ALLEN: Nineteen percent high, thirty-eight percent moderate, nineteen percent

low, twenty-five percent insufficient.

CO-CHAIR PARTRIDGE: Okay, moving on to validity. Any discussion? David.

MEMBER CELLA: This might seem like in the weeds a little bit, but it kind of gets to the NQF position on not endorsing measures, I mean, not endorsing instruments but endorsing measures.

And so my questions really are about A, the choice of a specific distance as opposed to a percent improvement. Have they looked at that because, you know, I've done some stuff in this area and this is a, if you can make a good case for somebody who might start at 250 meters, 25 meters may be a meaningful improvement.

But if they start at 500, that's proportionally only half the improvement and it's probably not all that meaningful. And they're going with a straight 25, and I realize that's been what's been used and you can maybe deal with the error.

But it seems to me, you know, thinking

about this, and I admit I've used this more in clinical trials and not in the real world, but a percent improvement would make more sense.

And this figure of 25 meters is low. It's the lowest figure of the debate in the area which ranges from 25 to 80. And I would have thought with individual classification which is what this is that you would go with a higher number, something more like 50, particularly in pulmonary rehab where people do pretty well generally.

So I guess those are my main questions, and then I have one other.

MR. LICHTMAN: Okay, those are good questions. Number one, you know, we went strictly by evidence based here. We didn't want to speculate, we didn't want to diverge from the evidence and the evidence basically says we look at the minimal important difference not a percentage.

One of the problems with a percentage is the patients who do better initially tend to

improve less than the patients who do poorly. Twenty-five meters was selected because this is a very disabled population.

When you do work with this population clinically, these are really very low level patients. And setting the bar too high would do an injustice because a 25 meter increase, we see that, and this is a little bit anecdotal, we do see that in our patients that when they improve by more than that, they really are feeling a lot better.

We had this discussion on our committee level and this was years ago when we first developed this, and it was decided to go with the 25 meters rather than coming up with a percentage improvement that we haven't verified in peer reviewed literature or going with a higher level.

Twenty five meters does appear in more of the studies than the higher level. And in the research literature that we looked at, particularly Holland et al., that's where we

derived it from.

You had to arrive at a cut point somewhere, and that was our rationale behind it.

CO-CHAIR PARTRIDGE: But to clarify, as I understand this measure, you start with where I was coming in. So I might be, say, at 40 percent functionality, David. But did I go up an additional X is the way I interpreted the specs. Am I right?

MR. LICHTMAN: Correct. Yes, that's correct.

CO-CHAIR PARTRIDGE: Yes, so it's not just across the board did everybody achieve 25 meters. You start with where I am and evaluate. Okay.

MEMBER CELLA: And the other, just a follow up, could I, just a quick -- so related to this and I'm sort of pushing the percent again with this. But more to the NQF than, you know, if this is a reasonable thing for this particular measure in this particular area, this situation.

But six minutes is arbitrary. It's

used because it's historical. There's nothing magical about it. The NIH toolbox is now a two minute walk test. There's good reason to think that you could do this in two minutes and not six minutes.

And so when you think about the migrating of a measure like this to other areas where there may not be as much willingness to take six minutes, you know, in this particular pulmonary rehab setting, if you had a percent benefit which does seem, I think, clinically reasonable, you have easier migration to say a two minute walk test or other performance tests that are even shorter to do because the goal here is to demonstrate a benefit in performance and not more meters specifically.

So I just make that as a recommendation that there be some way over time to move this toward percent benefit as opposed to a specific number.

CO-CHAIR PARTRIDGE: Thank you. And actually, things that we make as recommendations

can indeed be part of our formal report. So it's not just recommending to NQF. We can say it.

MR. LICHTMAN: And not as part of this report, we can take those recommendations and explore them without putting them officially into the report because I think that's a very provocative suggestion.

And we would have to do careful data analysis from our pulmonary rehab database to establish those cut points, and that's going to require more work in the future, which is fine.

And I think it's a great suggestion, but I wouldn't -- and to respectfully disagree at this moment, I wouldn't change from the evidence base at this moment. But we would certainly consider that in the future.

As far as the two minute test goes, running a clinical program for 20 years, the six minute walk test, it's not onerous on the staff. It was rated very highly as to the feasibility and the usability by the expert panel.

In looking at program certification

process, that's really not an issue in our programs. So most programs, all programs right now utilize a six minute walk.

CO-CHAIR PARTRIDGE: Okay.

MR. LICHTMAN: Again, if that morphs in the future, we'd be open to changing that.

CO-CHAIR PARTRIDGE: Okay. Validity voting. Nadine?

MS. ALLEN: Voting on validity, one high, two moderate, three low, four insufficient. Voting starts now. Eighteen percent high, fifty-nine percent moderate, eighteen percent low, six percent insufficient.

CO-CHAIR PARTRIDGE: Okay. Moving now, feasibility. Comments on feasibility? David, Karen, anybody? Ready to vote? Okay.

MS. ALLEN: Voting on feasibility, one high, two moderate, three low, four insufficient. Voting starts now. Thirty-five percent high, sixty-five percent moderate, zero percent low, zero percent insufficient.

Voting on usability in use, one high,

two moderate, three low, four insufficient information. Voting starts now. Forty-one percent high, fifty-three percent moderate, six percent low, zero percent insufficient.

Voting on overall suitability for endorsement of measure 0701 functional capacity in COPD patients before and after pulmonary rehabilitation, one yes, two no. Voting starts now. Ninety four percent yes, six percent no.

CO-CHAIR PARTRIDGE: Okay. It's now time for public comment from anybody in the room or on the phone.

OPERATOR: Okay, at this time if you would like to make a public comment, please press star then the number one. There are no public comments from the phone line.

CO-CHAIR PARTRIDGE: All right. Then it's time to move on and take stock of where we are. It's 5 o'clock. Our adjournment is scheduled for 5:15. I don't have the feeling that anybody wants to go further.

I think we've all had a pretty

difficult time today working through a lot of difficult measures and more to come. I am concerned we had one, two, three, four, five, six measures scheduled for 3:00 to 5:00 this afternoon, and we didn't reach any of them.

I do think that because of the similarity of five of them, probably we can shorten the time allowed from two hours. I have no idea since we haven't dealt yet at all with the measures that are generated through the Uniform Data System for Medical Rehab what the issues are going to be there.

So I'm going to turn to my colleagues on the right and ask them if they think that we are likely to achieve finishing this measure set tomorrow by 3:00 or whether --

MS. SAMPSEL: So our hope was that we could huddle with you and Chris. And I think we still want to do that. But we do have some similar groupings of measures for tomorrow that we might be able to have discussions with the developers about kind of re-working the agenda a

little bit.

But I know we also have a couple folks that need to leave early, and I believe, David, you need to leave by 10:30 at the latest? Okay, so we want to kind of figure those out a little bit.

I mean, I don't know if anybody has, you know, kind of aptitude or interest in spending the next 15 minutes or so talking about either of the ending questions which would help with tomorrow.

But we do also have the time set up. We have two hours on everybody's calendar next week to discuss anything we don't get to. So I think it's more of a is everybody done for the day? Do you want to have a 15 minute discussion?

I don't think we'll make it through another measure today.

CO-CHAIR PARTRIDGE: David?

MEMBER CELLA: I don't know. I mean, most of us are here. We could take the list of five. They may go very, very quickly. The five

improvement ones? The five CMS ones? They have a lot of similarity. You know, with the photo this morning, we did one and that really covered, you know, six or seven.

CO-CHAIR PARTRIDGE: Right. And they're at --

MEMBER CELLA: Then we wouldn't feel so far behind. But you know, that's --

MEMBER MONROE: You know, especially if we're going to lose people like David --

CO-CHAIR PARTRIDGE: We are.

MEMBER MONROE: -- I would much rather hear --

CO-CHAIR PARTRIDGE: My only caution is this is at a provider level we haven't talked about. It's home health.

MEMBER MONROE: Right, I was going to say because we're going to lose some people, I would rather hear about what they have to say on the two discussion points for the next few minutes, PRO-PMs for specific disease states and how we handle multiple conditions because we

won't, like, David won't be here tomorrow when we talk about that and I guess --

CO-CHAIR PARTRIDGE: I agree.

MEMBER MONROE: -- if other people have to leave early.

CO-CHAIR PARTRIDGE: And we're going to lose Chris as well because --

MEMBER MONROE: Yes.

CO-CHAIR PARTRIDGE: So, oh dear, all right. Okay. Then of the two, which one would you like to hear about, Helen, Sarah, Suzanne, which would you rather hear us talk about briefly first?

MS. SAMPSEL: I actually think we should talk about the second one because I think we're going to have some issues with that tomorrow.

CO-CHAIR PARTRIDGE: Okay. Discussion of parsimony and need for multiple experience of care in functional status measures for different settings? Who would like to open? Anybody given any thought to this from the agenda item? Chris?

CO-CHAIR STILLE: Well, I'll start. This is not by any means sophisticated or based on a whole lot of experience. But you know, I think to the extent possible where measurement techniques or specific measure sets can be used in multiple settings, they should be.

And to that extent, you know, approving different things for small variations in care settings, unless there's a good reason not to, maybe we should think about that being a default.

It would make things a lot easier administratively for, you know, groups that are administering the measures and figuring out what to do with them. And I imagine a lot of other people sort of share my feeling.

MEMBER THOMAS: Sorry, could someone just frame the discussion a little better? I don't really know where we're going.

CO-CHAIR STILLE: Right. So I mean, my interpretation of it is that if you have a bunch of measures of care that are relatively

similar to one another, that using them to measure small things like, you know, like differences in joints or differences in care settings between long term acute care and long term rehab care, to the extent that those settings and those organ systems or whatever are similar enough that they can use the same metrics, we should probably push for that.

MEMBER CELLA: You framed it earlier in your own words, and then I asked you to send me that document. There are a lot of measures, there are a lot of ways to get numbers, and sometimes they don't maybe need to be so diverse. And that's what you may want to talk about, right.

CO-CHAIR STILLE: I guess the critical question that should be asked of anything, you know, new is could you use what's already there and if not, why not?

CO-CHAIR PARTRIDGE: And from a historical perspective, NQF was founded in part to try to reduce the proliferation of measures

that looked somewhat alike and maybe were in most dimensions and to move toward, and we heard this theme from CMS all through today, move toward having a more standardized set of tools or of measures that everybody will understand and frankly, from the consumer perspective, that can be explained pretty readily.

And once you get used to how this is measured, you're trying to make a choice, provider or treatment, it's very helpful that oh yes, I've got that framework.

MEMBER LOEB: What you are saying is kind of limited. And I agree because you get to a point where your choices are so overwhelming.

I know this is just a really dumbed down comparison but when you go to the grocery store and there's 50 different tubes of toothpaste and you're just like I don't know what I need because there's just so many choices.

And that's what's happening because there's just so many different measures to choose from and there's going to be less people on each

measure to really measure because there's so many.

MEMBER NEUWIRTH: Can I just ask a point of clarification? Is part of the framing here that it's not just across settings but it's also across, you know, conditions or bodily parts and stuff like that?

CO-CHAIR STILLE: Just when I was looking at the body part things, it would have been nice to say, for the developers to say and this is needed because this is different and this is why. And that wasn't a requirement so they didn't do it.

MEMBER THOMAS: The one thing that comes out to me is I did say earlier and I still believe that there is this proliferation of measures and there's little gradations and differences between measures that we're looking at in different settings.

And it does seem as though a more standardized approach would be beneficial. The flip side of that is that there are certain

measures in certain settings that are well ingrained, that providers have completely invested in and follow and track.

And, you know, CMS has bought into and payment systems are designed around them. And basically care is delivered in a sense around meeting certain measures, almost like working toward the test.

So that's a pretty big, you know, disruptive thing to choose one over the other without even considering that I think. So I'm not sure that's part of our purview or whether we're supposed to be specific --

CO-CHAIR PARTRIDGE: No, it is part of our purview. We've had this discussion at the CSAC level, and Ann will remember some of it. One of the things, particularly with measures that have been around a long time and are good and are well ingrained but we've close to topped out, most everybody's performing it pretty well, we call and sort of reserve status or parking lot.

This is a good measure. It's out there, it's still a valid measure. We don't think that it's a cutting edge measure anymore. But for all kinds of reasons, your internal QI or something else, go ahead and use it, it's validated and useful.

MEMBER BRADLEY: I was just going to say, kind of working in a provider environment, this is an important discussion because it costs a lot of money to collect these measures on our end.

And not just in terms of human resources, but now we're in electronic medical records, and to retool an electronic medical record so that we can collect a similar measure but not exactly the same measure, it's very expensive and it just drives up the cost of care.

So I do think it's important to look at this as to why is it needed and what are the resources going to be required to collect these measures.

MEMBER BIERNER: One thing I would

like to add specifically coming from a rehabilitation background is we're moving more toward helping, working with the patient to establish what is important to him or her.

And I haven't seen a lot of that in any of the measures really that have come through today. But there are some things that we know won't change much.

I mean, in the area of rehabilitation some people may not make changes in certain areas. But there's not a lot here where we're soliciting input from either the family, the spouse, the caregiver in some cases, or the patient about what are your goals and choosing measures that are specific to them.

And so I think if there's any movement in that direction, rather than just having long laundry list of, you know, body part specific, we need to move more toward the patient when they're able, when they're cognitively able or the proxy for the patient, the caregiver, spouse, whoever, moving towards establishing goals at the onset

that could be measured with the help of the rehab team or others.

MEMBER KAPLAN: I'll be really brief, but I think the standardization versus the interpretation issue and the all purpose measure, if you think about math, and my husband hates it when I do this but it's intuitive for people.

It's like, if you think of a uniform math test that was going to do the whole population's math ability, right, and now I'm going to try and discriminate students in MIT one from another versus I'm going to use it in seniors on high school and I'm going to discriminate their performance one from another.

It's going to get trashed because it's going to have absolutely, no matter what you put out there, it's going to have absolutely no use at one of those extremes at all.

It's not going to vary. Everybody's going to flunk in high school and if you use the high school one, everybody's going to pass at MIT. So the tensions between standardization and

the unique applications for specific purposes often lead you to well, I've got this core set of things that kind of work generally, and then I have to modularize around specific applications.

So maybe instead of these either/or tradeoffs, we could think about. But then I'm now thinking oh my God, we're going to get into vendor wars about instead of different developers we're going to end up with the vendors being able to I can name that tune and further, you know, and faster and faster.

But if you look at the application under what am I trying to do, what's the purpose of measurement and is this appropriate for that purpose for that population, you're going to get safer than if I tried to do a completely standardized approach, I'm going to use SF-36 period, and I'm going to use it in every setting for all populations for functional status assessment.

MEMBER MORT: On the issue of specific versus general PRO-PMs, I think you can't have

your cake and eat it too. You just need some condition specific or disease specific measures, otherwise you'd be looking at something too generalized.

I think that's along the lines of what Sherrie was saying. And as far as parsimony or convergence or alignment, I would feel better about striving for that if there was a gold standard.

So rather than sort of be black and white about it, my point of view might be if there were questions that we could ask developers about, you know, have you tried to do it in a shorter form, have you piloted whether this format works better than that format, is there something similar.

So use it as considerations and actually tee up answers to the questions as developers bring forward their applications rather than say, you know, if there was a gold standard way of doing this that was simple, aligned, short, yes. But otherwise, I'm

struggling with it.

CO-CHAIR STILLE: Sherrie, did you have -- oh, okay. Dave, and then I'll talk.

MEMBER CELLA: I think I would like to offer a different perspective than what Sherrie and Liz just articulated which is that, you know, in that math example that you gave, Sherrie, if you have an item bank of math questions that runs the full gambit from very easy questions to very difficult questions, you can have your cake and eat it too.

That's what, I don't know if we're supposed to talk about previous applications, but that's what photo does. They have item banks and PROMIS does that.

And so it is possible. It's been demonstrated to have that cake. Now there may be differential item function by, you know, whether it's a knee or a shoulder or something like that in physical function and that's empirical question and can be tested.

So that is case by case, but that

vision is a realistic and reasonable vision to put forward that there can be one metric, one standard.

If you take, you know, let's take something that in some ways is a little simpler than physical function because physical function can be upper lower extremity, joint specific.

Depression, we've linked to one metric, the promise metric, you know, three other depression instruments including the PHQ-9. And I've talked here, actually part of the white paper that we put together about the concept and illustration, and now published illustration since that presentation last year of lining these measures up to the point where, and I think this is where NQF sits, you can have a performance measure that says you need to move a certain number of people up above this bar.

That bar may be set today by the PHQ-9, but because of this linking that works really well, you can replace PHQ-9 with PROMIS or with the Beck Depression Inventory or the CESD so that

you could come even more indifferent to whether somebody uses any one of those four things because you're concerned about the bar the same way you don't care which blood pressure cuff somebody used.

You want to know their blood pressure, you don't care which scale, who made the scale they're weighing somebody on, you want to know their weight.

So it is a vision that can be done for things like depression, certainly, I'm pretty certain at least, others may not be. Physical function is going to be a little trickier. There are the links that exist right now, and those are also published.

And all of this is pretty new. So you know, NQF and CMS find themselves now with this 300 page document that has all these different performance measures and people have to go out and load their EMRs with different things.

But the future could be on some of these things that are fairly generic in their

human relevance, depression, physical function, pain, there could be a common metric because underlying all this there is a common metric. And that I think is a realistic striving.

Oh, the other thing, I just to say, that's part of why in a related way I push on the edge a little bit on distance of meters walked versus percent change because if you start thinking in terms of a percent improvement, then it might not matter whether you're doing six minutes or two minutes and it might not matter whether you're doing a get up and go test or a six minute walk in various clinical applications because what you're caring about at the quality level is are you demonstrating a percent benefit to an individual patient, whatever that percent might be.

CO-CHAIR PARTRIDGE: Anne?

MEMBER MONROE: There is an appeal to me to, I hesitate to use the word common core, but that idea with extra questions put on given the circumstances or whatever.

And I'm wondering if the staff or anyone has looked across these similar measures to see if that common set already exists because it's in every one of these measures because when we think about having to develop one, it seems so, you know, so huge and full of argument.

But I would bet that if we looked across all these measures that we looked at today and probably tomorrow, there is a common core set that could get adopted.

And then extra two questions if it's your knee or an extra three questions if it's, you know, a nursing home setting. I don't know. But it seems to me there's got to be a way to get away from all of these things which should be a more common approach with the details being special to the circumstance.

CO-CHAIR STILLE: I would like to, just for a second, knit what Liz and Sam were saying into sort of a question for NQF folks is where the person and family centered care steering committee?

How much of a role do we have in advocating for stuff we'd like to see in measures that's person and family centered? So it may be, you know, up to us to in getting a common list we'd like to see these common things in the measures that we see.

You know, my experience as a researcher is that we're developing measures of coordination of care. And if there's not a parent input place, it's not as useful of a measure, for example. So how do we do that?

DR. BURSTIN: That's a great question. We've actually started part of the reason to move to standing committees was to have a body available to do exactly that.

But it shouldn't just be about looking at the measures that come before you. We actually just drafted a charter just yesterday actually to move towards what standing committees we hope will do beyond the measure endorsement piece which may include, for example, quarterly touch bases or specific projects, or also

providing the subject matter expertise for other initiatives like the MAP process for example.

We often have the chairs of the committee provide input to that process, as well. So I think that is very fair game. We would love to see help not only evaluate the measures before us but kind of move the whole field forward. So most definitely.

MEMBER THOMAS: What I don't understand about that, I totally get the approach, the value of doing something where, all right, now we've got all these measures before us, we understand them, we understand the evidence base.

But maybe we come up with this uber super measure that you can use in all these areas. But there's so much work and investment and time that went into preparing these things and these very separate measures, and demonstrating them and validating them and testing them.

How do you do that in a way that a

committee like ours wouldn't completely pick apart that work product? See what I'm getting at? I mean, maybe CMS does it because they're in a position to mandate it for payment purposes and then collects data and comes back.

I don't know. It just strikes me as being a pretty daunting proposition.

DR. BURSTIN: I think that was part of the basis behind the work CMS did around the CARE tool was to try to get closer to that vision of something that cuts across settings.

MEMBER BRADLEY: Well, I think there may be common measures that cut across a lot of settings. But I think they don't necessarily speak to quality in each of those settings.

Things like, for instance, right now we're reporting wounds. But that doesn't really speak to the quality of an inpatient rehabilitation hospital.

It might speak to quality of an acute care nurse, nursing unit, or a different type of setting. But I mean, it's something that we can

certainly measure, but it's not necessarily an indication of the quality and the mission of the setting.

So I think as you try to identify common measures across all settings, you have to be very careful that those measures are also measuring quality, not just that you can collect the data.

MEMBER MORT: I was thinking about David's comment about PROMIS and how it's a measure bank. And when you start talking about measure banks, you're really in my mind talking more about vehicles for dissemination and implementation almost.

So to me I'm thinking maybe this isn't a measure question as more as an implementation, a measure question as much as it's an implementation question.

So as we think about more and more patients in integrated delivery systems, patients at our medical homes, they're connected to their providers through portals or through electronics.

And do we want to think about harmonization and alignment more from the delivery of the measure and the measure collection rather than the actual specific indicator or metric because you'd go a long way I think towards improving health if you made it easier for the measures to be collected.

So rather than shoot for the perfect measure, shoot for more harmony around implementation. I'm sort of, I've moved to that way of thinking about it.

MEMBER LOEB: Chris, I want to just applaud you on what you said. What brought me to this committee was my experience and what Jared, my husband who passed away said. And you know, he was one of the measurement gurus.

And he said, you know, for 18 years I worked on measures. And he said once I became a patient, I realized that none of those measures had any impact on my absolute care and my patient-centered companionate care.

And I actually gave a talk to American

College of Physician Executives, and I was paired with a measurement expert. I'm like that's a really weird paring.

But it turned out it was a dynamic talk because he spoke of all the measures they had developed, and my presentation was loving and losing, is current measurement really the answer.

And so what you've suggested and what I think as a committee in addition to your endorsing the measures is to really sit down and see how we could bring person and, you know, family centered care away from just the hard measures.

CO-CHAIR STILLE: Right. But at the same time, we have the fire power, as it were, to be methodologically really, really good.

MEMBER LOEB: Right.

CO-CHAIR STILLE: And that's going to be the hard part, I think.

MEMBER KAPLAN: I think what Sherrie was getting at was the expanding the domain of observables, if you will, to include things that

could be sampled, a collective of which actually represents patient's perspective, the clinician's perspective, you know, the system's perspective.

That if we figure out a semi-permeable membrane that actually puts things in and takes them out depending on the purpose of measurement, et cetera, et cetera, then we've got something that's really more meaningful to all levels of --

MEMBER LOEB: Because it's like, I mean, oh, I'm sorry. I mean, I know HCAHPS is a big thing. And you know, we can't do away with it. But truly, HCAHPS doesn't measure whether it was really a good experience for the patient.

You know, yes was your patient experience good and how you rate it. But when you really drill down to it, when you're in a hospital and getting prolonged care, it's not picked up by that. So we really need to get to the patient or the person as it's now called.

MEMBER NEUWIRTH: Yes, I would echo, Sherrie, exactly what you're saying. I think that one of the things it feels like it's missing

in both our process but also with some of the developers is really what's the patient experience of all these measures?

I am pretty confident that the majority of patients have nothing to do with the outcome of these measures. They don't know where they go, they don't track them, they don't see them.

And maybe it does impact their care, but I think there's, you know, I think that there's lots of variability in how actionable a lot of these measures are.

My mom was in a skilled nursing facility, had a five star rating. And I got to see five star rating care close up. And it was so, so horrible.

They did not detect C. diff, I had to, like, you know, really go all out and really plea to have the doctor come. And she ended up getting rushed to the emergency room and dying of complications from C. diff that they didn't really detect early enough. And there were many

signs and symptoms of it.

So you know, what do these measures mean? And I think that of course we need measures. But with the overwhelming number of them, it's dizzying both for us as providers and health plans and advocates, and it's ending up to be, you know, somewhat not that meaningful for patients.

So I guess I would really appreciate more feedback from patients. And maybe there needs to be, like, a deeper kind of study of the sort of comprehensiveness of the measures really starting with patients and families and doing some deeper inquiry into what would matter to them.

You know, I hate to bring up the Yelp! app, but every one of us, when we need to go out to a restaurant, how many of us go to Yelp!? You know, and it's simple, it's easy, it's there. And it's a simple rating. It's, like, totally simple.

And you know, I'd love to go Yelp! and

see a skilled nursing facility and be able to trust it. Now how much do we trust Yelp!, I don't know. But you know, it's not that far off, and it's continuing to get better.

But I do think that there's, you know, something in between this sort of overwhelming, what feels like overwhelming number of measures we have right now, and instruments. And it's really hard to implement.

I mean, we have had challenges implementing the PHQ-9. And One Group Health, one of our affiliate partners has done a terrific job. They're, like, leading and they're implementing what 30, 40 percent of the time.

They're leading in the country for PHQ-9 use. And that's, like, a measure that's been around a really long time. And that is a powerful measure.

So you know, I'm sorry I'm ranting here but I just feel like I think that it's been a little, I think, frustrating to see the number, the proliferation of these measures and not

really know how they're being implemented and used.

And I guess I would, I think the parsimony sounds like a great idea. I also think that to get to parsimony we need to have input from patients and families about the sort of comprehensive.

Let's have them look holistically at this because, you know, I'm sure that some of them, some of us because we're all going to be patients too and we all are at various times patients.

But you know, if we've had knee surgery, we do want a question about our knee, right? Or something related to our knee. But does it have to be completely, radically different instrument and set of questions?

So I'll stop there. Sorry for the rant. But I have a feeling many of us feel the same way about this, that we could be doing much better.

MEMBER MORRISE: You know, one way we

can involve patients and families in an area where patients and families consistently have expressed all the patients and families I talked to across the country concern is that they give input in a measure or they're asked in HCAPS whatever.

And sometimes somebody will even say would you like me to get back to you about your concern? No one ever does. And so I think particularly one thing we could do on measures, particularly if there's a patient reported outcome, when we get down to use, we could look to see if it's being used to counsel in any way with the patient or share information and data with the patient.

Let's get back to them. They took the time to say this is what's going on. In use, do we see that the patient then is engaged in the end process? That would be one thing.

MEMBER DOWDING: Yes, I guess I've been struck by today and also by reading the stuff for tomorrow how absent patients are from

anything in the documentation that we're looking at.

With a patient, person centered outcomes measurement committee, there's not even a question on the NQF assessment about whether or not patients have been involved in --

(Off microphone comment)

MEMBER DOWDING: There is, but it's not filled in. It's not something --

(Off microphone comment)

MEMBER DOWDING: Yes, and it might be there. But it's not something that's highlighted as it being important enough. It's not one of the criteria that we vote on, for instance.

It's not something to do with, in the importance is it a specific thing that we say has this been identified by patients as important, have they specifically said it's important, have patients been involved in developing this?

Where's the effort and the space for it? So it's sort of there, but it's something that's skirted over, it's not something that's

part of the procedures that we use.

And I just wonder if we might need to have slightly more discussion about, especially when you've got so many measures that are looking at the same thing.

I'm sure function is important, but I'm not entirely sure I've picked up from any of the measures today that the patients have been asked if it's important.

CO-CHAIR PARTRIDGE: I hope this has been helpful to you all. I think it's been very helpful to me anyway. And I encourage all of us to copy Karen -- who's the cake here -- who tends to ask the developer the extent to which patients were involved in some of the testing.

It's a good -- pardon?

(Off microphone comment)

CO-CHAIR PARTRIDGE: I'm sorry, Katherine, yes. It's a good thing to put the developers on notice that we're going to ask. Okay. I'm in favor of saying we've had it for the day.

And Chris and I will huddle briefly with the staff. And we hope to see those of you who can still stay, we promise not to talk about performance measures over dinner. But please do come and join us.

(Whereupon, the above-entitled matter was concluded at 5:30 p.m.)

W

C E R T I F I C A T E

This is to certify that the foregoing transcript

In the matter of: Person- and Family-Centered Care
Phase 2 Standing Committee Meeting

Before: NQF

Date: 01-21-15

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