NATIONAL QUALITY FORUM

+ + + + +

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

+ + + + +

THURSDAY
JANUARY 22, 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
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:

NADINE ALLEN
HELEN BURSTIN
MITRA GHAZINOUR
ANN HAMMERSMITH
MARCIA WILSON

ALSO PRESENT:

SOPHIA AUTREY *
JEROME CONNOLLY
KEZIAH COOK *
BETH DEMAKOS
ANNE DEUTSCH
DANIEL DEUTSCHER
DAVID GIFFORD
CHRISTINE GOERTZ *
JAYNE HART CHAMBERS
DAVID HITTLE *
BEN JOHNSTON
NICOLE KEANE *
MARJORIE KING *
TRACY KLINE *
STEVE LICHTMAN *
JASMINE LARSON *
JANE LUCAS *
STACY MANDL
TARA McMULLEN
PAULETTE NIEWCZYK *
POONAM PARDASANEY
COLLETTE PITZEN *
LINDA RESNIK *
GARY REZEK *
ANGELA RICHARD *
JEANNETTE SHRIFT *
LAURA SMITH
ANITA SOMPLASKY *
ELLEN STRUNK

MARK WERNEKE

TRACY ZHENG

*** Present via telephone**

TABLE OF CONTENTS

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

8:34 a.m.

CO-CHAIR PARTRIDGE: Good morning everybody. I know we're still getting caffeinated and all that early morning fog is still clearing a bit, at least for me.

But, we want to be able to move along pretty expeditiously this morning. We are losing people. We've already lost, I think permanently, Carin van Zyl, who's ill and she indicated she would try to phone in from her hotel room. But, frankly, I'm hoping that she's sleeping and getting better before she has to fly back to California.

David is heading off to NIH to make a speech and I know that Deb and Esther, I believe, are both slated to leave at noon, or maybe Esther is not. You're okay.

Okay, I've had this in my head that we had a couple of people with planes that were the -- that they had to leave by noon and Chris has to leave for his plane by 2:00.

So, we may lose our quorum at noon, in which case, I think almost certainly we're losing our quorum at noon, in which case we will continue with our discussion of the measure, we will just not vote in person on them. And if we need to discuss then still further on our call, we'll make this judgment at the end of today.

If we need to discuss them further on our scheduled call next week, we will. If we don't, we'll send a SurveyMonkey survey out fairly promptly after this meeting and we'll vote by email.

We will probably also not attempt to do the discussion of related and competing measures. However, I do want to call your attention to the part of the package that was in front of us is a cheat sheet that Sarah put together to help us as we think about related and competing.

I recommend not losing it because she's laid out quite nicely the difference in a chart form, the differences between the various

measures. And I personally find it very useful as I'm trying to keep in my head what I'm considering as related or competing.

CO-CHAIR STILLE: That'll be your homework for the trip home is to review that.

CO-CHAIR PARTRIDGE: Right. So with that, I'm going to turn the gavel over to Chris who's going to take us through the first set of measures and we'll do housekeeping.

MS. SAMPSEL: So, we do have just a couple of things.

One, I know for some folks, there were some issues at the hotel with how they were doing charges. Some people they put on the master bill, some people they charged your credit card. They have somebody's suitcase in hostage. And we'll try to -- you know, we have our meeting department working on that to figure that out.

But everything was supposed to be direct billed, so if your credit card was charged, we ask that you watch it, notify us because we'll try to get all -- the goal is to

get all of those charges reversed.

I somehow got a new scarf last night. It showed up on my chair at the restaurant and I don't think I have kleptomania issues. But if this is anybody's scarf, it started living with me yesterday and I apologize because I didn't intend to take it. And it's nice, it's cashmere, it's made in Germany.

And then, you know, I think, Liz, I don't know if you're prepared, I think you had a couple of questions.

But I think the other thing we just wanted to do was kind of regroup a little bit and I want to remind you all that you do have the decision logics, algorithms in front of you on how to work through each measure based on what is presented regarding importance in evidence and then through scientific acceptability.

You know, I think we want to make sure that if anybody has questions on how they should be voting based on information presented that you ask that before we vote. But, you know, and we

don't want to spend a whole lot of time kind of rehashing criteria, but if you do have that need, we can certainly do that.

We also just wanted to revisit where we were yesterday or what we ended up with yesterday and out of the 28 measures, we did make it through 13. So, I think, you know, there's some success there. We got into the double-digits.

And the first slide, just, you know, as a recap with the FOTO measures, so 0422 through 0428, those were the measures that are technically in NQF-speak are currently not recommended because they failed at the importance criterion and the developers will have the opportunity to provide more information prior to the end of public comments. So, we will re-discuss these measures.

We did not -- we only got to, this is I think the highlight of the day, 0688 is, you know, went through as recommended. The other measures we will be discussing today, actually

those will be later this morning.

And then for long-term care hospitals and some of the outpatient rehab for long-term care, 2631, which was the percent of long-term care hospital patients with admission and discharge functional assessment and a care plan, that measure was not recommended. It did not pass the importance criterion.

With outpatient medical, with the outpatient measures, we made it through all of those where we had one recommended, one not recommended and two in the gray zone.

The gray zone measures do move forward as recommended but we will re-discuss those as well.

So, that's where we were with those and if anybody has any questions regarding process and kind of overall criterion adjustments, if we could do those now before we start heading into the measure discussions.

MEMBER MORT: I just had a question about, I felt, and maybe if I read these

algorithms in a more detailed way I would understand how we did what we did.

But, I thought at the end when we recommended the pre- and post- six-minute walk test for pulmonary patients that it passed as an endorsed measure. But what they recommended, what they were proposing wasn't the pre- and post-, they're proposing a percent change across patients.

So, I thought we sort of gave them a by and we went ahead and approved it anyway. And I felt that we had been a little bit more strict in looking at exactly what the recommended measure was and critiqued it based on exactly what the developer was proposing earlier on.

Am I the only one who felt that way? And I think in some -- if I am, then --

MS. SAMPSEL: I think that's what I'm just trying to look around and see if there's folks who are.

MEMBER KAPLAN: What bothered me a little bit was the shift from the morning to the

afternoon and it was some part of sort of a direction that -- and maybe this is not what you're getting at, Liz, but I thought we kind of were shifting into we're going to look at, you know --

Usually, I understood that NQF only reacts to what it sees, not what it thinks it could see if other analyses were done or different changes were made or, you know, different assumptions were held or something.

And, Sarah, I kind of felt that same shift from morning to afternoon that we were kind of maybe doing things differently in the afternoon than we have done in the morning.

MS. SAMPSEL: Well, I guess my question then would be -- I mean there were differences in the measures. So, in the morning we really were, I would say the vast majority of the measures were patient reported performance measures, the PRO-PMs and in that case, the NQF standard is higher than just the outcomes and the process measures.

So, I don't know if that was part of the shift that you were seeing that there really is a difference in the criteria based on those measures and I guess, you know, as staff, we can, you know, keep that under consideration of, you know, how do we make sure that we separate those out well so that there's an understanding of the shift.

CO-CHAIR STILLE: I think that was a particular issue with, I think it might have been the last one we talked about because they were proposing to use it for accountability at the organization level and they had only tested it at the patient level and we had a lot of problems with that.

MEMBER MORT: That's exactly the one that got me a little muddled at the end. And I think in part it's because these patient-reported performance measures are new.

Having been on the group that worked through the concepts a couple of years ago, it's difficult to sort of get through it

methodologically and understand exactly what a PM is or a PRO-PM is. So, I can understand the developers are working through that as well.

So, I guess what I would ask is, today, if the chairs and facilitators would just be really clear as to remind us and help coach us along as to exactly what measure is being proposed so we have the right construct in our head and then we can move through it in a more reliable way I guess is what I would say rather than shifting a little bit.

MS. SAMPSEL: Okay. And then I think the other thing that, you know, regarding, I think that as your 0701, what we can do with that is, you know, and this with all of the measures, is we have the -- we'll write the report, there's the public comment period and then there's a post-public comment call where you always have the opportunity to say, this isn't what we were, you know, we approved or this is a totally different concept and we can reopen that vote.

MEMBER KAPLAN: Can I ask one other --

and not to slow us down, but can I ask one other clarification question?

Some of what I've been looking at and reading the reliability and validity sections in these measures, they use the reliability at the patient level and they're very careful to document how reliable it is at the patient level.

But then the proposed use is at the facility level and there is no evidence of reliability at the facility level.

So, is NQF, what's the instruction to us on that issue?

MS. SAMPSEL: And we tried to get at that a little bit yesterday. And so, when you follow through the algorithm what it does is that you almost to the bottom.

And so, the first criteria is did they perform reliability or validity at that measure of population level that they're reporting at? And if the answer is no, you go down another level and did they do it at the patient level or item level if it's off an instrument?

And then if you follow that across, you get to the point of you can vote it as moderate or low.

So, frankly, the criterion is it's an and/or criteria. You either do it at the measure level and the patient level or you do it at the other level but you have a lower grade of, you know, you would give them a lower rating when we vote.

MEMBER KAPLAN: So and now, why wouldn't you give it an insufficient? Because if there's no evidence, I mean if there's one level of evidence but the level of evidence that they're proposing to use the measure at is not there, why wouldn't you give it an insufficient data? Because you don't know if it's going to be low until they run the runs.

MS. SAMPSEL: And I'm not sure I'm really the best person to answer that one. But I think technically you could, as a committee member, give it insufficient. So, your choices there at the bottom, you could do that.

But I would also say that, you know, I think even Helen admitted that, you know, the kind of difference in criterion between if it's process outcome or the patient-reported outcome is something that I think NQF needs to revisit.

And especially for these measures. And I think that's bringing -- these measures are bringing that to light that that really is an issue.

CO-CHAIR STILLE: And Sherrie, you would say insufficient because the developers might actually have those data that they can then bring back and present, yes.

MEMBER KAPLAN: Yes, I mean you want them to get -- if they have the data or they have an opportunity to get you the data, you'd certainly want to have it rather than give it a low because it's already been done and it's low, I would feel much more comfortable --

CO-CHAIR STILLE: That makes a lot of sense, yes.

MEMBER KAPLAN: -- saying it's

missing, you know.

MEMBER MORT: And I just have one more question about goals.

I was under the impression that when we were approving measures for NQF endorsement, the construct was these measures would be used for accountability purposes and not for QI purposes.

And if that's changed, and that's what I thought I heard peppered throughout the conversation yesterday, I just want to have clarity around that because that would change my whole way of thinking about the voting if we were approving them for QI.

Oh, here's Helen.

And if the criteria for acceptance or endorsement include both acceptability as an accountability measure or suitability as an accountability measure or a QI measure, that makes me think about it slightly differently.

So, could we get clarity on what we're endorsing, Dr. Burstin?

DR. BURSTIN: Yes, Dr. Mort, yes. It's, again, a perennial issue for us. At this point, we do have an expectation that a measure that's endorsed by NQF is available for all purposes. That could include, you know, QI.

I think what we don't have are measures purely for QI. And I think there's always an expectation they can fulfill the full range of accountability applications as well.

As I mentioned yesterday, that's in play as we move to potentially moving towards endorsement by ratings or grades or intended use.

MEMBER MORT: So, it has to be good enough for accountability, but of course, it could be used for anything?

DR. BURSTIN: Exactly.

MEMBER MORT: Is that the way we think about it?

DR. BURSTIN: Yes.

MEMBER MORT: Got it.

DR. BURSTIN: Good.

MEMBER MORT: Thank you.

CO-CHAIR PARTRIDGE: And ideally, accountability and public-reporting, but obviously, in a number of instances the public-reporting is planned as down the road, it hasn't been used that way yet.

MEMBER MORT: Thanks for the clarification.

CO-CHAIR STILLE: Okay, great. Let's begin our long task this morning.

We're going to start, I don't know if we mentioned, the order is going to be pretty different to both facilitate speedy reporting and speedy analysis and also help people that need to go to different places.

We're going to start with the UDSMR measures 2287, 2286 and 2321.

We're then going to do the American Health Care Association measures 2613 and 2612.

After the AHCA, we'll do the CMS measures 0167, 0174, 0175, 0176 and 0177.

And then the IRF ones from CMS 2635, 2633, 2634, 2636 and 2632.

And hopefully, we'll get through most, if not all, of those. If we -- we may well lose our quorum and if that happens, we'll have discussions but not voting right now. So that's going to be the plan.

MS. KEANE: Hello?

CO-CHAIR STILLE: Hello?

MS. KEANE: Hi, this is Nicole Keane from Abt Associates. We're the contractor for the five measures from Home Health that were at the end Day 1.

Usually when the day goes over, we start with the measures that were meant to be done on Day 1.

And our developers are shifting their schedules so that they can be available this morning. We are going to lose people at 10:30.

CO-CHAIR STILLE: Okay, and which measures are those?

MS. KEANE: 0167.

CO-CHAIR STILLE: Okay, yes. We anticipate --

MS. KEANE: So, I would respectfully request that we could --

CO-CHAIR STILLE: Right, right, we still need to do UDSMR first. We will -- I believe we'll be able to get those done this morning, but so we may do some shifting depending on how long the UDSMR measures take. Does that make sense?

MS. KEANE: Thank you.

CO-CHAIR STILLE: Great. Great.

Okay, so we'll start with 2287, Functional Change in Motor Score at IRFs and I believe Sam and -- no 2287, Sam and Deb were the discussants for that and they're both here and that's good. So, okay, go ahead.

MS. DEMAKOS: I'd just like to confirm that Dr. Paulette Niewczyk is on the phone with me as well. Paulette, are you there?

MS. NIEWCZYK: Hello, I'm here.

MS. DEMAKOS: Okay, great.

So, first of all, I would just like to thank everyone. Our schedules are shifted a

little bit, that may be a good or a bad thing, depending on how early morning you're used to being.

But, I'd like to thank the committee for their work and for the opportunity to present our measures to you.

Just as a little background, in 1987 a task force of physicians, therapists and researchers was charged with establishing an instrument to measure functional outcomes, medical rehabilitation and patient burden of care.

After three years of research, measure development and instrument testing and validation, the FIM instrument was developed.

The FIM has been endorsed by the American Academy of Physical Medical Rehabilitation and the American Congress of Rehabilitation Medicine.

The work was a result of a federal grant awarded to the researchers at the Center for Functional Research, CFR, at the University

of Buffalo. Today, CFR remains a division of the U.B. Foundation and is housed within UDSMR. It's also a not-for-profit division of the U.B. Foundation whereby the subsequent maintenance of the FIM has been occurring for the past 25 years.

For those not familiar with the FIM, it's an 18 item measure that measures patient function and burden of care. It's currently used across the post-acute care continuum which speaks to the core measure set that was mentioned yesterday, and we do have subsets that are abbreviated versions for other venues.

Each item is rated on a scale from 1 to 18, I'm sorry, 1 to 8 7 which refers to -- one refers to complete dependence, seven which refers to complete independence and the overall range is 18-126.

It might help to bring up the measure form that we submitted with this so that people can see a visual.

We're thrilled that the functional measures are now being considered by the Patient

Family Centered CARE Committee.

Function affects every person and function is a high priority for patients in all venues of care. A greater level of function allows a greater level of independence for the patient and a decreased burden of care on family members in care settings.

The FIM has been used by thousands of clinicians from interdisciplinary rehabilitation teams for case management, monitoring and patient goal setting. And patients can be directly involved and we encourage that by participating in goal setting with their care givers.

The FIM provides estimates of patient burden of care. In other words, the number of hours the patient requires one-on-on assistance from another person for personal care on a daily basis in the home setting and community.

As the FIM ratings increase, the minutes of care per day decreases. The FIM contains several subsets of measures within, three of which we are putting forth today for

endorsement.

The proposed change in measures score, Measure Number 2287 that has been put forth by UDS is constructed by utilizing a subset of 12 items from the FIM, eating, grooming, dressing upper body, dressing lower body, toileting, bowel management, transfers to bed/wheelchair or chair, transfer to toilet, locomotion/walk or wheelchair, locomotion/stairs, expression and memory.

These items are currently collected in the IRF setting and imbedded in the IRF-PAI as the instrument developed by CMS and used for the IRFs to assess functional outcomes and for payment by Medicare. So the burden of collection is low as the IRFs have routinely collected these measures.

The change in self-care that we're putting forth as well which is Measure 2286 and the change in mobility which is 2321 are also subsets of the FIM and subsets of the form that we're putting forth today for mobility.

And I think I'd like to turn it over to Paulette to talk about the evidence.

CO-CHAIR STILLE: Just quickly, if any of you are following along, it's one of the Appendices in the SharePoint.

MS. NIEWCZYK: Hello? Can everybody hear me?

CO-CHAIR PARTRIDGE: Yes.

MS. NIEWCZYK: Okay. Thank you so much, Beth.

My name's Paulette Niewczyk and I'm sorry I couldn't physically be there in person today. So, I thank you very much for the opportunity to speak on the line.

And when we had submitted the measures to the NQF, we had a little bit of back and forth with some of NQF folks because there was a little bit of confusion and I can understand why that may be the case.

If you have taken a look, and I know there's a lot of material to review, you may have noticed that the motor score, or I'm sorry, that

the motor measure is in essence the self-care measure as well as the mobility measure.

So, it is important for us to give you a little bit of that background and history about the FIM. I assume several of you may be, you know, very well versed in it, but some may not.

The FIM is the larger entity, so it's 18 items and it's been used very widely in inpatient rehab as Beth had mentioned for several decades now. But it's not only been used in the inpatient rehab setting, skilled nursing facilities have used the tool as well as long-term care facilities.

So, it has been an instrument that has been used in many different post-acute care venues to look at patient outcomes and function as well as patient burden of care as was already mentioned.

These new measures that we're submitting are new in the sense that they're being looked at as a separate entity. But the items are not new.

So, I want to make that clarification because they are verbatim the same items that do exist in that larger FIM instrument.

So, and one of them that I mentioned, one of them does contain the self-care plus the mobilities. So, when we talk about motor, there's 12 items, but you'll see that the other two measures are nested within.

Now, why did we submit three separate measures?

Because depending on the patient for goal setting or depending on the venue, it may be more critical to look at self-care alone without being nested in that composite score, so that total summed score.

So, we felt it was important, especially for patients that may not ambulate, and it may not have to do with their course of care for a specific condition. So, if they haven't ambulated their entire life, we didn't feel if we're looking at quality outcomes that it would be important to necessarily fault the

facility or the clinicians on something that they may not be working on because it has nothing to do with their condition that they're presenting to.

If you'll notice the evidence, we reference the FIM extensively and that is because, as I said, the measures are new. We have not been testing them as their own entity, they are exact items that are currently living in this larger instrument.

So, the reliability and the validity have been extensively studied and I did provide bibliography with each of the measures, but I'd be happy to provide a more expanded bibliography on some of the reliability predictive validity, construct validity and so forth.

I do want to clarify that even though each of these measures, the items within are being rated on a 1-7 scale, all of them have a Rasch conversion. So, they can be used not only in their ordinal properties but also in a linear property.

So, we do have those item level raw to Rasch conversions and that will actually be a bit more sensitive to patient change. So, when you're looking and you're trying to compare the patient to themselves or between patients or between facilities, it allows you to do some of those facility level comparisons.

The data that we presented and provided within the submissions have been from actually a random sample from our large data repository. So, this is -- we have data about half a million patient level cases per year and that's just in the IRF venue.

So, we took a random sample to provide these cases. And, of course, you know, we could certainly expand that in any capacity.

And we had presented only the inpatient rehab facility results to you. So, we do have data repositories on skilled nursing facility as well as long-term care hospitals. That is not provided in our submission but if you would like additional analyses we certain could

provide that as well.

CO-CHAIR STILLE: Great, thank you.

MS. NIEWCZYK: I feel it important to mention that in terms of feasibility -- oh, sorry.

CO-CHAIR STILLE: I think we need to start the discussion. Everyone's kind of looking at me. Sorry.

MS. NIEWCZYK: Okay, sure.

CO-CHAIR STILLE: So, I just had a clarification question before I turn it over to the discussants.

So, all of the items on 2286, 2287 and 2321 are part of the FIM and the self-care items plus the mobility items make up the -- so 2286 plus 2321 equals 2287, am I right?

MS. NIEWCZYK: Correct, exactly.

CO-CHAIR STILLE: Okay, okay. So, that's important for people to understand. Good.

The other thing is I think just for purposes of doing what we need to do to figure out how to vote, we need to really start to talk

about importance first and then we can jump into the methodology.

Okay, so, Sam, are you going to start?

MEMBER BIERNER: So, this is Sam Bierner.

The use of the FIM primarily has been in the inpatient setting. And for those of you just so you understand, this is a clinician-derived score and requires training of the clinician who can be a therapist or a nurse or a physician has to undergo a training session to be certified to give these scores. So, it's a fairly detailed process to learn to do the FIM scoring.

So, as I read it, you're wanting this to be a one year score. This score is going to be scored one year after an event or after discharge from rehab? That's my question.

MS. NIEWCZYK: No, so the time period is one year. But the way that the patient would actually be scored is by looking at their rating, the change from their admission to their

discharge.

So, if a patient had stayed in an inpatient rehab for, let's say, 21 days, they would have an admission assessment taken within 24 hours of admission to the facility and then they would have another one done within 24 hours of discharge. And it would be that change from the admission to discharge.

MEMBER BIERNER: So, really -- okay, so, what is the purpose of saying it's a one year score? That threw me off in your writeup.

MS. DEMAKOS: I think that's for benchmarking purposes, Paulette.

MS. NIEWCZYK: Yes, that's exactly it.

So, in addition to looking at the patient level which is critical for that patient and that family, we also look at aggregate data comparing, you know, all patient and we adjust the adjustment methodology is included in the submissions, but also at the facility level.

And then you can look at some best practices or see what some of the variability is

over the course of a one year period.

MS. DEMAKOS: So, I would just like to clarify one thing as far as the training and credentialing as well.

UDS, I mean if you were to use this instrument through us and request benchmarks and outcomes, you would be required to do the training and credentialing.

However, CMS is using this same instrument within the IRF-PAI and they do not require training or credentialing to use that instrument.

So, there's --

MS. NIEWCZYK: Yes, thank you. That's an excellent point, Beth.

So, there is extensive guides that are available to in essence self-teach. So, you could certainly -- we have those available and you can certainly do the assessment without the extensive credentialing and passing on the master exam.

However, it's been demonstrated

through a number of studies that in order to display the very high rater reliabilities, it's critical to have some basic understanding of how to do the rating, otherwise it's just -- there's vast inconsistencies.

MEMBER BIERNER: Right, and from personally using frequently, I can tell you that the training is necessary. We've had to do a lot to ramp up training of others in our facility to get better inter-rater reliability.

But, what about the patient that goes through inpatient rehab and then six months later has had additional therapies as an outpatient? Are you expecting that you'll do another score then? Otherwise, you're just using it during the inpatient rehab state. So, are you expecting you'll use it at other times on the outpatient basis?

MS. NIEWCZYK: That's a great question. On the outpatient side, the FIM certainly can be used as an assessment instrument. However, it's key to look at the

construct that's being measured and what's being measured is, is patient function, but it's also burden of care.

And burden of care really refers to how much time that patient would require from a helper, another person, one-on-one if they were living within a community setting.

So, if the patient was now in a community setting receiving outpatient therapy, it's not to say that the patient's level of function doesn't require any additional services, that's not what we're saying by any means. However, they may not require an hour or more of helper assistance.

So, in essence, you're going to see a ceiling effect or you're going to see patients top out and it's just because the instrument isn't sensitive enough to pick up on some of the very, very small but still important elements of function that a patient may need once they are able to, you know, community-dwell, go back to work and so forth.

So, at that point, it might just mean other additional measures would be used to check that patient.

MEMBER BIERNER: Okay. So, the --

MS. NIEWCZYK: I do want to say it's not just used for inpatient rehab. So, we do have, and have for the past 20 years, have patients assessed in the skilled nursing facility as well as in long-term care hospitals.

So, all of the items that we have submitted to NQF would be applicable and appropriate for those other levels of inpatient care.

CO-CHAIR STILLE: I'm sorry to cut you off again, we just need to be as brief as we can --

MS. NIEWCZYK: Oh, sure.

CO-CHAIR STILLE: -- in assessing these things.

MEMBER BIERNER: So, the instrument itself has been used a long time and has a large database, so I don't have a problem with the

instrument. I'm not quite clear on the distinction of the three different scores and how you're calling one a motor score when it encompasses -- it's a composite score of many different functions.

But other than that, the instrument itself, you know, I think has certainly been validated and has reasonably generated reliability when it's used with trained clinicians.

CO-CHAIR STILLE: Other thoughts about importance, gap in care or gap demonstrated up to for improvement disparities?

MEMBER SALIBA: I did not see the data on disparities. I may have just missed it in here. Was there information about disparities?

CO-CHAIR STILLE: Not as far as I know.

MS. NIEWCZYK: No, thank you for asking that.

So, we weren't exactly clear as to how in depth the analyses should be and we absolutely

can and we'd be very thrilled to provide that information to you.

We do have patient level as well as facility level data on different patient characteristics, so we certainly can stratify and look at these outcomes by race or by payer source or by age category.

And I think it's critical, especially on the heels of last night's discussion, I was on the call I think until nearly the end, and in looking at all payers I think is an important thing.

So, this is not restricted only to a Medicare/Medicaid population or to an age 65 or older. We do have data on 18 and above and would recommend its use among all adults ages 18 and over, all payers, all races and socio-economic strata, status, strata and so forth.

So, we do have that available and we can provide it to NQF.

MEMBER SALIBA: An additional follow-up question.

I know that you indicated that it is done in multiple settings and that's true, the physical therapist in SNFs will often use the FIM because they've been trained in it, but it is not required in SNFs or nursing homes.

Are you targeting this measure towards particular types of facilities? Because I noted in the feasibility section you're talking about it being, you know, collected as part of the IRF-PAI. Are you proposing this as an inpatient rehab facility measure or are you proposing -- so, what type of setting or institution are you proposing this measure for?

MS. NIEWCZYK: We propose this to be used in inpatient rehab, skilled nursing facilities, long-term care facilities and as well as home health facilities.

So, all of this inpatient or quasi-inpatient as I like to refer to them, facilities could benefit from using these three measures that we have submitted. All items would be applicable to patients in those venues.

So a ceiling effect as I spoke to a couple of minutes ago with an outpatient venue wouldn't be applicable with those four that I had just mentioned.

MEMBER SALIBA: So, when you looked at the --

MS. NIEWCZYK: The data submitted were only on inpatient rehab --

MEMBER SALIBA: Yes, when you looked at the --

MS. NIEWCZYK: -- because that is where it is used in terms of the IRF-PAI, however, as I mentioned before, we do have data in those other venues and we could share that with the committee.

MEMBER SALIBA: Okay. So, I guess I'm trying to understand the proposal that you've put in front of us now would be for this to be a measure in all settings or just to be a measure in IRF?

MS. NIEWCZYK: In all settings. So, all settings except in outpatient. So, home

health, long-term care, skilled nursing and inpatient rehab.

Again, I guess I'm going reference yesterday's conversation late afternoon again, somebody mentioned, I believe it might have been Ann Monroe, but somebody had mentioned the importance to have a tool or a measure that could be appropriate in, I think she referred to it as a common core, in all post-acute care venues.

I think that's critical and that's really what we're posing here. We have items that have been well tested and validated and used in payment for the IRF facilities but --

MEMBER SALIBA: Yes, I think --

MS. NIEWCZYK: -- voluntarily by some other facilities.

MEMBER SALIBA: Yes, can I ask --

CO-CHAIR STILLE: I'm sorry, we're going to have to cut you off again.

MEMBER SALIBA: Yes, so the only other point would be that the data that's broken out here right now is not broken out by settings,

correct?

MS. NIEWCZYK: Correct.

MEMBER SALIBA: Okay, all right.

MS. NIEWCZYK: This is Just IRF.

MEMBER SALIBA: Right, this is just IRF data which is where we would expect the performance to be probably the best.

MS. SAMPSEL: Yes.

MEMBER SALIBA: Okay.

MS. SAMPSEL: Yes, so a couple of things. One, you know, the committee will be asked to review this measure and vote on this measure as it's presented with the data presented which is the inpatient rehab facility.

The other thing is I just want to, you know, we're really, again, getting out of control on time, so we'll be asking as we go forward, you know, if the chair's asked the developers to respond then at that point you can respond. Otherwise, developers should not be responding at this time.

And I think otherwise, are we ready?

CO-CHAIR STILLE: Yes, Peter?

MEMBER THOMAS: A quick question. So, I am glad to see that -- well, clarify for me if you would, I read expression and memory are part of what is recorded in the self-care measure, is that correct?

MS. DEMAKOS: Correct.

MEMBER THOMAS: So, that's -- I just want to make a point that that's critical because you, you know, the whole cognitive development of some of the patients that are in this setting is really critical and you're sure you can move to some degree, but you don't know where you're going if you don't have the cognition. That's obviously a big factor.

I also wanted to say that if you're looking at just inpatient rehab, I can see these measures being appropriate. I'm a little concerned that outside of the inpatient rehab setting that do these measures go far enough?

Are they fairly elementary in terms of the ability of the person to really be fully

mobile as opposed to -- I see locomotion and stair climbing, but is that it for locomotion? I'm Just trying to get a sense for --

I can see in the inpatient rehab setting that's a big -- those are two big important measures. But, if you have any further information about additional measures on mobility in particular that would round out the experience or the assessment of someone's real ability to be mobile and ambulatory?

MS. NIEWCZYK: Yes, so for the mobility measure there is --

CO-CHAIR PARTRIDGE: Paula, sorry.

Peter, again, I just have to go back. We're judging on what's in front of us and that's kind of a -- if you were going to use this another setting question is where you're heading and I think --

MEMBER THOMAS: Okay.

CO-CHAIR PARTRIDGE: -- maybe we should not take time to push it.

CO-CHAIR STILLE: Any other questions

on importance before we vote on importance for this measure?

Becky?

MEMBER BRADLEY: I just had a comment that I think it is important that the measure be applicable to more than just Medicare patients because we have more than Medicare patients in the inpatient rehab facilities. And so, it is a measure that can be used across all payers.

CO-CHAIR STILLE: Okay. David?

MEMBER CELLA: Just a proposal that this vote be carried over to the other two so we'll do three at once because they're very overlapping.

CO-CHAIR STILLE: Right, I think we might want to do -- should we maybe go all the through this one and then consider the other two kind of quickly and see if the votes for the first will carry to the second? Okay, then the third. Great, thanks.

Okay. You're proposing that we vote, yes, okay, great. She's got the clicker in hand.

Okay, so voting on importance for Measure 2287. Nadine, tell us when.

MS. ALLEN: Voting on evidence, one second, the vote. Please do not begin voting until I say start the votes. That's why I had to go back.

Now we're voting on evidence, one yes, two no. Voting starts now.

MS. SAMPSEL: I think you're missing one.

MS. ALLEN: Oh, we have 17 now. Sorry, guys, it's saying we've got 17 votes but I'm not seeing the votes.

CO-CHAIR STILLE: All right, should we just do a hand vote?

So, yes?

(A SHOW OF HANDS)

CO-CHAIR STILLE: Is that everybody? Okay, 17 yes, zero no.

MS. ALLEN: Let's try this again. Voting on performance gap, one high, two moderate, three low, four insufficient. Voting

starts now.

All votes are in, 24 percent high, 47 percent moderate, zero percent low, 29 percent insufficient.

Voting on high priority, one high, two moderate, three low, four insufficient. Voting starts now.

All votes are in, 53 percent high, 47 percent moderate, zero percent low, zero percent insufficient.

CO-CHAIR STILLE: Okay, let's have a brief discussion on reliability and validity.

MEMBER BIERNER: Just as I mentioned before, the reliability is very dependent on a trained clinician. It will certainly suffer reliability by someone who isn't trained in the use of it. It's not as easy to use as some measures that we've looked, but when trained, it can have high inter-rated reliability.

And it's validity has been well established I think over the last 20 years or so.

CO-CHAIR STILLE: Sherrie?

MEMBER KAPLAN: Yes, can I ask the developer a question?

Okay, so especially if this is going to come up again and again, there is a lot of evidence obviously for reliability at the patient level and the Rasch modeling that you did looks like there's only -- there may be one clunk item in this.

But that kind of modeling suggests that, you know, the reliability at the patient level certainly is good. But, I'm concerned that the reliability at the agency level, you've got a beta binomial model here and the interclass correlation coefficients, to me, look like a measure level mean variance were used to estimate rates as opposed to the composite score which is what I think you're going to use within dimensions to evaluate the performance of these institutions.

So, it doesn't look to me like you actually rated the interclass correlation coefficients at the facility level to compare in

the unit that -- the groups of measures you're actually going to compare.

You did it at the measures level it says on page 33 of your analysis as opposed to the composite across scores by agencies.

Is that accurate?

MS. NIEWCZYK: Yes, that's correct. We just had only so much space at the end to that application would allow us to include. But we have that other data and we can make it available. But you are correct in your interpretation.

MEMBER KAPLAN: So, you could make those data available to us?

MS. NIEWCZYK: Between facilities? Absolutely.

CO-CHAIR STILLE: And, David?

MEMBER CELLA: So, we're talking about the total FIM measure now, but the next two will be subsets of that. So, this question kind of relates to all three.

And I mean my first question comes out

of a concern that I've actually in my own work been grappling with for ten years now which is, you know, it's just by way of quick background, when you convert raw sum scores in to Rasch or IRT scores, you convert them from ordinal measures to integral measures so that the distance between each number is equal no matter where you are on the scale. And that's nice, it sounds nice.

But what it actually does functionally, and I think this to me is very important to understand here, is that inevitably when you're in the middle of a distribution, it takes more raw score change units to get the same interval change as opposed to the tails.

So, therefore, if you're using the black-box Rasch-transformed score and you have somebody that the extreme end, a very small change will look very big on the Rasch interval measure.

And I still don't know after thinking about this for years which one's correct.

So, my question is, I mean I tend to suspect the raw score is more correct because I'm not convinced that the same amount of change at the extreme means more than in the middle, but it's really a belief, I don't know.

So, my question is, since they have so much data, this is to the developers, are all of these determinations and scores based on the Rasch-transformed score? And if so, have you ever compared that to the raw score changes to see if that would change the individual or facility level change?

MS. DEMAKOS: Paulette, can I give it the quick --

MS. NIEWCZYK: Great question.

MS. DEMAKOS: Paulette?

MS. NIEWCZYK: But this is a project we're working on aside from this. But, yes, the answer is yes.

We have looked at raw to Rasch conversion and we've kept it in its raw or its ordinal level state and we're working on a

publication to submit. We didn't provide but the data you're seeing are Rasch-converted.

So, it's patient level where their 1-7 have been converted to what their Rasch-transformed value would be.

Yes, you do see bigger jumps so from the 1 to a 2, whether it's in its raw or Rasch-converted state is a big jump as well as the 6 to 7.

By converting to a Rasch measure, it helps to correct for some of those drastic differences. It lets you know that they're not all to be assumed equal. A 2 to a 3 is maybe not the same as a 1 to 2 would be. But that also adds to its clinical significance as clinicians would probably know that it's probably easier to get a patient from a 2 to a 3 than it is from a 1 to 2 in practice.

So, it would -- the Rasch-converted value would give you greater sensitivity at that patient level and likely mimic what you'd see in practice with treating a patient in practice.

MS. DEMAKOS: Can I just add quickly here, too, that we've also taken the raw scores and taken those and converted those to minutes of care so you know what level, how much time it takes for the burden of care with those raw numbers as well.

CO-CHAIR STILLE: Okay. Lee has a question then Sherrie.

CO-CHAIR PARTRIDGE: I just want to understand the risk-adjustment methodology a little bit because I think it's different from what we've seen in some other measures.

As I understand going through the steps, you'd start by classifying your patient in to one of the impairment groups and you calculate the patient score. And then you take a look at what I call a facility case mix.

In other words, sometimes you're looking at the patient and doing the adjustment at the patient level and then saying now we're going to measure the change beginning where that patient was.

As I understand it in this process, you're doing that but you're also looking at the facility and stepping back and saying, what does the facility's case mix look like?

And I'm calculating what is, in essence, an adjustment at the facility level so that if 80 percent of the patients were this type and 20 were that type, I then adjust the results not based on the patient's -- you see why I'm getting all tangled up?

MS. DEMAKOS: Yes, I do. Yes, I do see what you're saying. So you're saying --

MS. NIEWCZYK: Yes, you can -- we do adjust at both the patient level as well as the facility level. That allows for comparisons between facilities.

So when we're looking at, you know, actually quality or outcomes by facility, if you were to speak with a facility, they're always going to say, oh, my patients are different. My patients are more severe. My patients are worse.

Not only would you be able to look at

that severity, but you can also then compare facilities that are truly like the -- meaning if it is a designated, you know, stroke center, you can compare that facility to other facilities that are also designated stroke centers or --

MS. DEMAKOS: But I think what you're -- Paulette, I think what the question was really is that you're looking to compare patient to patient within a facility, is that what you're saying?

CO-CHAIR PARTRIDGE: As I understand this --

MS. NIEWCZYK: You can do that as well.

CO-CHAIR PARTRIDGE: Wait a second.

MS. NIEWCZYK: So, you can get at the patient level but you can also do it at that facility level so that way, you're really seeing apples to apples comparison, correct.

MEMBER KAPLAN: I just wanted to follow-up on the earlier question about some of these items in the Rasch scores.

Could you give us, if you're going to give us the ICCs at the facility level, could you also give us the mean square fit statistics that you use, you know, in-fit and out-fit?

MS. NIEWCZYK: Yes, absolutely.

MEMBER KAPLAN: Because that would help answer that question I think.

MS. NIEWCZYK: Absolutely.

CO-CHAIR STILLE: Okay, anything else about reliability and validity?

Dave?

MEMBER CELLA: It's just -- I want to make sure I'm clear that we're voting on the use of now the overall motor FIM in inpatient rehab facilities using the Rasch score with an understanding that they do have raw score data to compare and they're working on academic work to -- is that --

So we're voting on the Rasch score.

MS. NIEWCZYK: You're voting on the Rasch score, yes.

All of the data has -- it's collected

using the 1-7. So, when the clinician does an assessment, they don't do anything with Rasch, they don't use the Rasch-converted value. They only use a 1-7.

On the back end, in order to validate these measures, we performed Rasch analysis and we converted each of those values. So, a 1 then became converted to a Rasch-transformed.

So, I'll give you an example. One might be a --

MEMBER CELLA: I'm sorry -- I understand all that. I was just clarifying what we're voting on. I know it's all been --

MS. NIEWCZYK: Yes, it is.

MEMBER CELLA: Okay, thank you.

MS. NIEWCZYK: Transformed, yes.

CO-CHAIR STILLE: So, this is just sort of a question -- a procedural question, so, it sounds like there's a lot of data people are interested in. We're voting on the data that we have. After the vote, depending on the vote, does that determine whether or not we need to get

more data or can we request that or sort of what happens with that?

MS. SAMPSEL: We can still request it before public comment.

CO-CHAIR STILLE: Okay, before public -- okay, great.

MS. ALLEN: Voting on reliability, one high, two moderate, three low, four insufficient. Voting starts now.

One more vote. Thank you.

Thirty-five percent high, 35 percent moderate, six percent low, 24 percent insufficient.

CO-CHAIR STILLE: Okay.

MS. ALLEN: Voting on validity, one high, two moderate, three low, four insufficient. Voting starts now.

Twenty-four percent high, 53 percent moderate, zero percent low, 24 percent insufficient.

CO-CHAIR STILLE: Okay, any discussion on feasibility?

I have just a quick question on feasibility. It appears as though the public availability and use of this is dependent on kind of how you're going to use it.

For the purposes of accountability and reporting between institutions and benchmarking, is that all within the public domain? I was a little bit confused as to the description.

MS. DEMAKOS: The benchmarking piece of that is not available to the public. I mean as far as that's concerned, you would go to a vendor if you wanted benchmarking. So, but this would be made available to the public, obviously, but the purposes of use for a patient level.

MS. NIEWCZYK: Measures used, yes.

CO-CHAIR STILLE: Okay.

CO-CHAIR PARTRIDGE: But, going back to Sam's earlier question, as I understand it, it would be -- there could be some problems if your staff isn't trained in using FIM.

MS. DEMAKOS: We totally promote training and credentialing with this instrument,

not everybody does it, but we would highly suggest it and agree with Sam's comment.

MS. NIEWCZYK: And there's other models that could be used, like a train the trainer. So, it doesn't mean you have train all clinicians within an entire facility. You may select one or two and those then could go back to the facility and train the others.

CO-CHAIR STILLE: Dave?

MEMBER CELLA: I'm just going to repeat myself here, but I think it's important that, and particularly with feasibility, that this vote is an inpatient rehab facility vote, it's not other sites. Because there was some comment earlier where there was a recommendation for other inpatient-like sites. But this is an inpatient facility vote on feasibility.

MEMBER BIERNER: It's not exactly what she said, though. She said she's going to use it in others -- I mean because skilled nursing -- other settings can maybe used in as well. So, it's not as well - those are not, in our view,

those are not inpatient settings even though they're classified here.

MEMBER SALIBA: But I think the measure is just that we're voting on that's in front of us, it's just for IRFs. That's what I was trying to clarify earlier.

MS. NIEWCZYK: In the submissions we indicated on each of them the settings and we listed all of the four venues that I spoke, IRF, SNF and LTCH and home health.

MEMBER SALIBA: So, I'm incorrect in my interpretation that this is a measure just IRF that you've put in front of us?

CO-CHAIR STILLE: Yes, you're incorrect about that.

MEMBER SALIBA: Okay.

DR. BURSTIN: But we only endorse measures for the levels for which we have testing. So, if we've only been provided data on the IRFs, then that is the only setting applicable today.

MEMBER SALIBA: So, I'm a little

confused about what we're voting on. I'm sorry.

So, are we voting on this as a measure for any setting or are we voting on this as an IRF measure?

CO-CHAIR STILLE: An IRF.

MEMBER SALIBA: IRF, okay, great. Thank you.

MS. DEMAKOS: May I ask a question?

So, if we've submitted the data just for the IRFs, would it be possible then, again, to submit it for the skilled nursing and the LTCH to be considered for that as well?

UNKNOWN PARTICIPANT: Not today.

MS. DEMAKOS: Oh, no, not today.

MS. SAMPSEL: I mean I think we'll have to think about that internally. But I mean the way that this was submitted and, you know, clearly, even in the measure title were you have IRF, we were considering this as an IRF measure and, therefore, as one measure submission.

If, you know, if you want to submit additional data, I almost think that's a new call

for measures.

CO-CHAIR STILLE: Okay. Oh, Liz?

MEMBER MORT: I have just a quick question.

Since these data are used for IRF-PAI with regard to payment, does Medicare do any auditing for quality on sites?

That might just mitigate the concerns around variability and implementation somewhat.

MEMBER BIERNER: I can't -- the facilities can be audited, yes. There's RACs which audit the IRF facilities.

CO-CHAIR STILLE: Anything else on the feasibility issue or should we vote? Let's vote.

MS. ALLEN: Voting on feasibility, one high, two moderate, three low, four insufficient. Voting starts now.

All votes are in, 18 percent high, 65 percent moderate, 18 percent low, zero percent insufficient.

CO-CHAIR STILLE: And usability? Any usability comments? Okay, let's vote.

MS. ALLEN: Voting on usability in use, one high, two moderate, three low, four insufficient information. Voting starts now.

Thirty-five percent high, 53 percent moderate, zero percent low, 12 percent insufficient.

CO-CHAIR STILLE: And then finally, suitability for endorsement. Any other last comments before we vote? Okay, let's vote.

MS. ALLEN: Overall suitability for endorsement of Measure 2287 functional change.

Change in mode of score, one yes, two no. Voting starts now.

All votes are in, 83 percent yes, 12 percent no. Sorry, 88 percent yes, 12 percent no.

CO-CHAIR STILLE: Okay, so there's been a lot of discussion that crosses the UDSMR measures which is great.

What we've decided to do is allow the discussants here on the panel any opportunity for making more comments on 2286 and then 2321 and

then we'll figure out how to vote.

Dave?

MEMBER CELLA: I just have one thing that's, you know, specific to really these two subset, or I guess you could say short forms. But their sub-domain related short forms, and it relates to the reliability data that were provided.

The coefficients are kind of moderate and normally, I would just assume that they are adjusted for item overlap but I didn't see that in the report.

So, when you correlate a four item mobility score with a 12 or 18 item overall total score that includes those four items, you're going to inflate that relationship because four units of information are identical.

So, were those correlations that you reported on reliability adjusted for item overlap or do they include the common items in both sides of the equation?

MS. NIEWCZYK: The inter-item

correlations were specific to that measure. So, for the mobility, they were how well each of those four items correlate with each other in the measure. So, it was not independently correlated to the 18 items.

MEMBER CELLA: There's a section though where you talk about the correlation of the four item measure, for example, with the total. That's the one I'm referring to, not the inter-item.

MS. NIEWCZYK: Okay. So, what we did is taking those four items, how well do those four items and only those four items predict what the patient's full 18 item FIM score would be. So, that we ran regression analyses to see how well those individual four items of that measure could predict the full 18 item value for that patient.

MEMBER CELLA: All right. Well, it would be useful, yes, I mean just as a comment then, given the magnitude of the coefficients which are around .6, that suggests to me that

that's a relationship between the four items that are subbed out and the other items, it's fairly low which ironically may be a case for these two sub-measures as better measures than the overall.

But that's just what -- I think it would be worth knowing the adjusted correlations and you might want to consider over time migrating to these two sub-scores as better than the overall since the overall seems a little cloudier.

CO-CHAIR STILLE: Interesting. Sherrie?

MEMBER KAPLAN: When you reduce item batteries like that, that drastically you take a hit in Cronbach's alpha because of the -- well, what I'm staring at is.

And so, you would expect that in Spearman-Brown that that's almost exactly what you'd expect.

So, my concern, however, was as with the other measure, you don't have -- you don't do it at the facility -- the inter-class correlation

coefficients are missing for facility level reliability, is that accurate?

MS. NIEWCZYK: Correct. So, this all just done on the whole sample of patients, not at the facility. Now, we do have other data that was not submitted and that's not in front of you. So, this is at that patient level.

MEMBER CELLA: Okay.

MS. NIEWCZYK: And thank you for also stating that. Yes, so the fewer items you have you will take a hit in terms of the reliability, the more items, typically, the more reliable the measure.

MEMBER CELLA: I need to just clarify, it's in your validity section, actually, because you're citing it as validity and I'm not talking about the internal consistency or the inter-time correlations within. I'm talking about the relationship of the mobility score with the total FIM and the relationship of the self-care score with the total FIM.

Those coefficients are moderate,

they're not high. And given --

MS. NIEWCZYK: That's actually not just correlations, though. We're looking at the proportion of variance accounted for. So, a point --

MEMBER CELLA: That's true, you've got --

MS. NIEWCZYK: -- that's pretty respected if we know roughly 60 percent of variance could be accounted by those four times alone. That's very respectable.

MEMBER CELLA: Yes, that's -- but not when you have common items. You're explaining variance of one side with the same thing on the other side. So, I think you should separate them out.

But my point really is not against the measure, it looks okay and I just wanted to make that clear. It's that I think you actually -- the data suggest over time, you might be better off with the two sub-scores as more valid than the overall. And in some cases, more -- and in

some sense, more reliable. But, that's all.

Thank you.

MEMBER KAPLAN: Can I make on point of clarification?

If you square the correlation -- the inter-items, the reliability coefficient, that's the maximum reliable variance that you can share with another variable.

So, if you square the lower reliability coefficient, you get a lower number that's reliable variance to be shared with other variables.

So, just for clarification, if you shrink the score, you're going to take a hit in reliability and that in turn is going to compromise your ability to see reliable variation in a different variable.

MEMBER CELLA: Okay.

CO-CHAIR STILLE: so, the other discussant, let's see, Brian was the other discussant for 2286 but he is not here.

Sam, did you have any other specific

things to 2286 or are you pretty much done?

MEMBER BIERNER: I'm done.

CO-CHAIR STILLE: Okay. And then, 2321 was Becky and Dave. Dave, you've made a few comments and Becky, did you have anything else you wanted to say about this particular measure?

MEMBER BRADLEY: I just wanted to comment on the question about accountability and auditing and just kind of reiterate, there is a lot of auditing and oversight by CMS related to the payment side of these measures. So, that is occurring.

CO-CHAIR STILLE: Okay. Ann?

MEMBER MONROE: I'd like to ask Sam a question.

You mentioned several times the need to be well trained in delivering this instrument. Because you've raised it so often, do you have concerns that the lack of trained people will impact the strength of the results in terms of how well this tool is being used?

MEMBER BIERNER: Yes, because we know

from our own internal data looking at different groups that we have to train -- residents and we have to train new therapists and people that without adequate training the inter-rater reliability is very poor. No, I won't say very poor but it's not very good.

And so, it definitely requires training. That's why I was concerned about them talking about using it in LTCHs, SNFs and other settings where they aren't as well trained or may not be as well trained as in an acute or rehab hospital.

So, it definitely requires a lot of training. It's not as easy to administer as some instruments are.

CO-CHAIR STILLE: Okay. So, should we take separate full votes on each of those quickly?

(SIMULTANEOUS SPEAKING)

CO-CHAIR STILLE: Yes, all right. Okay, so should we take a vote on assuming the votes from the prior one applying to these? Any

objections at all?

(NO AUDIBLE RESPONSE)

CO-CHAIR STILLE: Okay. Okay, so noted. Great, thanks.

CO-CHAIR PARTRIDGE: So, in response to the concerns about whether we have our developer resources available, we're going to rearrange again and move to the Home Health Measures which begin with 0167 and that's?

MS. SAMPSEL: So, that's the Abt folks on the phone. And do we have CMS folks here as well?

MS. KEANE: So, this is Nicole Keane --

MS. SAMPSEL: And before you start, I just want to be really clear, you really only have three to four minutes for any type of introductory statements.

If the Committee Members then have questions for clarification, we'll do that. But I really need you to restrict your opening comments.

MS. KEANE: So, this is Nicole Keane from Abt Associates. We have Tara McMullen from CMS in the room and then we also have fellow contractors from Acumen, Keziah Cook and then also from Colorado, David Hittle and Angela Richard who'll be presenting information.

Keziah?

MS. COOK: Yes, can everyone hear me?

CO-CHAIR PARTRIDGE: Yes.

MS. COOK: Great. So, I think one thing to keep in mind is these five Home Health Measures are all based on the OASIS instrument and they're all calculated and risk adjusted to getting some more methods and then, additionally, the data we present on reliability and validity are, you know, conceptually very similar.

So, there may be some efficiencies of considering some of these features across measures as your discussing results.

So, I'll start just with the introduction for 0167 which is Improvement in Ambulation or Locomotion.

Many patients who receive home health care are recovering from an illness or injury and have difficulty walking or moving around safely.

In particular, they may need help from either a person or from special equipment to ambulate or locomote.

Home health care staff can encourage the patient to be as independent as possible and can evaluate the patient's needs for equipment or other devices to help them move around.

And very importantly, improving a patient's safe ambulation and mobility are absolutely critical to allowing that patient to remain in their home rather than moving to a facility-based setting.

Even improving functional status, especially ambulation and locomotion contributes to the patient's quality of life and allows them to continue to live safely for as long as possible in their own environment.

So, overall, recovering independence and walking or otherwise moving around with

assistive devices is often a goal of rehabilitation provided in the home health setting. So, it makes it a reasonable evaluation indicator of effective and high-value home health care.

Overall, about 74 percent of home health patients are eligible for this measure. And the measure is calculated using the Home Health Outcome and Assessment Information test that's called OASIS.

There's a specific item on the OASIS that documents several levels of ambulation.

So, basically it starts at a level zero which indicates the patient has no impairment to their ambulation, moves through, you know, a couple of values that correlate to using a single-handed devise or a double-handed device to walk around.

There another level that indicates requiring assistance from a person for ambulation.

And then finally, there are several

categories indicating wheelchair bound but able to wheel independently, wheelchair bound and not able to wheel independently and then finally patients who are bed-fast.

So, the improvement in ambulation measure uses this item captured at two different time points. So, it uses the items that's measures when the patient first enters home health care at the start of care and that's based on the OASIS assessment conducted within 48 hours of the beginning of home care.

And then there's a second OASIS assessment that's conducted when the patient is discharged to the community setting. So, when home health care has specifically achieved their goals and is going to be no longer caring for the patient.

So, in both cases, the patient is assessed on this six point scale and if their numerical score on that scale decreases, which means they move from a sort of higher need category to a lower need category, then they're

considered to improved in ambulation.

The home health agencies as a whole are scored based on all home health quality episodes that end during a 12-month period. And the measure is their fraction of patients who improve in ambulation and then for public reporting purposes, that fraction is risk-adjusted based on patient structure.

So, I'll stop there.

CO-CHAIR PARTRIDGE: Thank you.

We have -- Carol Levine was scheduled to be one of the presenters on this one together with Peter. Carol is, unfortunately, not with us today because she is ill. So, Peter, you're up.

MEMBER THOMAS: I just have two main comments.

First is that I found the rationale to be very supportive of the need for rehabilitation in terms of home health care but less in terms of the value of measuring ambulation or mobility on the individual.

I recognize the value of being mobile

and in terms of living independently and performing manual tasks and all those things. But, I didn't see any data or any background about, you know, the say spinal cord injury patients who are non-ambulatory and lose bone density or muscle atrophy or the perils of lying in bed all day and, you know, contractures and bed sores and things of that nature.

So, I was wondering why that wasn't -- that seemed to me to be the obvious correlation between measuring mobility and ambulation and improving health care. Why was that omitted from the packet?

CO-CHAIR PARTRIDGE: Would our developers like to respond? I'm not getting a response.

MS. COOK: I can speak up.

So, you know, I think our focus here has been sort of across the range of ambulatory abilities.

Our fraction of patients in home health who are actually bed-bound, I don't have

the numbers in front of me, but it's quite low and, you know, I think it's under five percent.

And I guess the important thing to note is that if a patient is bed-bound at the beginning of the home health episode and then transitions to being able to even sit in a wheelchair, even if they can't move about independently in the wheelchair, that is captured as improvement.

I think another thing to keep in mind is that one of the core services provided by home health agencies are physical therapy. So, it is a care setting in which significant sort of physical rehabilitation is a goal.

We do have other measures that aren't in front of this committee that specifically get at things like pressure ulcers that are potentially more applicable directly to the small sub-population of the home health patients who are bed-fast.

CO-CHAIR PARTRIDGE: Thank you.

MEMBER THOMAS: Fine, well, let me go

to my main concern with the measure and that involves the Jimmo v. Sebelius case and settlement. And this is something that Brian raised yesterday. I just need to make sure people are well aware of this if you're not.

The CMS and the Department of Justice and the Center for Medical Advocacy and selected beneficiaries settled the case a few years ago called Jimmo v. Sebelius that determined that in the home health setting, the SNF setting and in the outpatient therapy setting, Medicare cannot impose an improvement standard that's different than IRF. IRF you have to have an expected improvement in the patient.

And so when we're measuring, I kept looking for how you accommodate for that decision because if this measure goes to eventually a pay-for-performance model, which I can assume all of these measures will eventually be looked at in that light, you're going to have significant incentives to not accept patients that require only maintenance or prevention of deterioration

kinds of therapy and don't have any expectation of improvement.

And if you're just measuring improvement in ambulation in this setting without accommodating for that, that's a major problem. That's a disincentive to treat those patients for home health agencies to accept those patients and provide services because they're going to get dinged for those patients.

So, my question is then, all right, well, how do you accommodate for that? Can you put that into the denominator as an exclusion or can you risk-adjust that?

If you do it on a -- I suppose you could do it on a condition-specific basis, but you always leave someone out. If you did it on a functional basis, then you're undercutting the whole purpose of the measure because you wouldn't want to keep out people that weren't improving, you're trying to measure improvement.

So, can you help me this? Because this runs not only throughout the home health

measure but the SNF measures and any outpatient measures and I don't see it really accommodated for in any of the materials that I've seen.

CO-CHAIR STILLE: So, it sounds like this would be a usability issue potentially.

MEMBER THOMAS: Well, it may be, but, boy, this is a critical for me and for a lot of the folks that I kind of deal with. This is critical because you don't want to have a measure that ultimately creates a disincentive to treat the very patients that you're trying to make sure are getting access to good quality care.

So, I would hope to elevate it above usability. I think this goes to the evidence issue.

MS. COOK: Is David still on the line from Colorado?

MR. HITTLE: Yes. Hello, this is David Hittle from the University of Colorado.

And the fact is that the preponderance of folks who are getting home care are relatively short stay in home care, around 35 to 40 days

generally and are generally those who do have a reasonable probability of achieving an improvement in their functional status.

Jimmo may have changed that somewhat in recent years. I don't have concrete data on exactly how much that would have changed, but we certainly are looking into that.

But we also have other measures actually related to functional status that look at whether or not you manage to keep somebody at the same level of impairment in functional status.

Generally, those measures tend to be very heavily topped out and, therefore, you know, usually during the first parts of a home care episode, is it, you know, there's not much differentiation in terms of whether or not you're able to keep somebody reasonably stabilized in their functional status.

And actually, we had, at one point, there were stabilization measures that were -- there were one or two stabilization measures that

were endorsed by NQF and NQF withdrew that endorsement several years ago because of the topped out nature of those measures.

We are currently exploring whether or not we can look at an overall functional status stabilization measure and whether that would, you know, if you look across several different ambulation, bed-transferring, bathing, dressing, the different kinds of ADL activities whether or not you would end up getting a greater level of variation, and therefore, a greater ability to distinguish among providers on those measures by like inspecting and, you know, constructing composite measures.

CO-CHAIR PARTRIDGE: Thank you, Dr. Hittle. I'm going to cut you off because I think we really need to explore other pieces here.

And Peter, I'm very sensitive to your concern about unintended consequences, so I don't want to terminate the discussion, I just have a feeling that it's probably going a little deeper than we can accommodate right now.

MEMBER THOMAS: Yes, and I don't know whether this is more in the validity, but I'll raise it under the validity part so we can move on.

CO-CHAIR PARTRIDGE: Okay. I would invite the people who looked at bathing, bed-transferring, management of oral meds and pain to also chime in here since the measures are somewhat similar. So, go ahead.

MEMBER PARISI: So, this is Len, and I reviewed the second one on grooming and bathing.

I am very familiar with the data collection process and the use of the OASIS. It's been in use for a long time and the rigors involved in educating staff to be able to perform a valid and reliable assessment and a lot of effort goes into that.

The only thing that I have a question about is the gap between the measured outcome and the evidence to support those interventions that would support improvement.

Much of the data that we collect related to these outcomes are used for performance improvement initiatives and benchmarking purposes. But the way the -- and the way it's reported on home health compare, the language is different for the public as it is on these measures as well which sometimes causes confusion among staff members.

But, I don't see any evidence that supports clinical initiatives or practice parameters around improving these outcomes. That would be my only comment.

CO-CHAIR PARTRIDGE: Do the developers want to respond? They've submitted it and we didn't find it or they would submit or?

MS. COOK: Is Angela on?

MS. RICHARD: Yes, I am. Hi, this is Angela.

You know, I think in the literature, we actually, particularly for bathing I guess, there are several studies that really indicate that home health care can really result in

improvements and functional ability, some were ever specific to bathing. The Luft study of 2009 particularly calls that one out.

And, you know, so I'm a little bit unclear about where the literature -- I mean I think the literature really does support this home visiting nurse intervention versus care as usual.

Friedman, Lovell and Powers 2014 found that and those are in the lit review. So, maybe I didn't make those clear enough.

MEMBER PARISI: Well, I guess maybe I wasn't clear.

So, the improvement in bathing, so if an organization scores poorly, there's no, to my understanding, connection between how you get to improve bathing.

So, is it education? There are no practice guidelines around educating people on bathing. It really gets back to their functional ability.

Clearly, it's an important indicator

because you want home health patients to be independent and to be able to have the ability to bathe themselves.

So, that was -- it's a little different in an ambulation where you would have a physical therapist and/or some of the other measures, so that was the only thing that I was trying to put together.

MS. RICHARD: Okay. And we did --

MS. MCMULLEN: This is Tara McMullen from CMS.

A lot of these measures, this kind of goes back to our discussion yesterday about the use of the nursing home measure and how they're used and commonly reported and benchmarked for those multiple purposes.

Many times we create these outcome based measures so that we can benchmark or set thresholds or publically report beyond that of the guidelines. We know guidelines exist and we take those very seriously.

You know, crafting process or crafting

how care is given, we know that these measures are a part of that but many times, like with this measure specifically or the other ADL measures, they're used as just basically, it's a mark, it's a benchmark and that's how they're reported on Home Health Compare.

So, they could be used to craft practice, but the means for those outcome measures at this time is just a benchmark and publically report what is going on within that setting.

CO-CHAIR PARTRIDGE: And I saw multiple other hands over here.

Lisa?

MEMBER MORRISE: Thank you.

I looked at bathing and transfers. The thoughts that I had around these, a couple of thoughts.

One is that I understand what Len is saying that there's -- when you determine that an organization does not meet a ceratin level of being able to provide these services, there's no

follow-up currently.

But that's a usability issue and one that we may want to look at if it's going to be used for payment so that there's some kind of motivation to improve. Because there currently is a gap between the small agencies and the large agencies in ability to provide those services.

I really appreciate what Peter had to say. I think, though, that that gets to a policy issue in also with usability versus a measure. The measure is still the measure and that we need to look at policy if there are unintended consequences to the measure when we get back to usability so that we have some kind of follow through. So, that's all kind of interconnected.

And I know that in our experience with home health interventions like these, we do reach a threshold where we've met goals and then home health goes away and then we fall backwards.

So, there needs to be communication with the patient so that there's kind of ongoing assessment.

But anyway, those are just my thoughts. I'd love to vote.

CO-CHAIR PARTRIDGE: Managing drugs and I'm --

CO-CHAIR STILLE: Yes, the last two are a little bit different from the others. Maybe --

CO-CHAIR PARTRIDGE: Yes, I think --

CO-CHAIR STILLE: -- Med Management and Pain.

CO-CHAIR PARTRIDGE: Pain's one of my favorites.

CO-CHAIR STILLE: SD had Med Management and Sherrie and Deb had pain.

CO-CHAIR PARTRIDGE: Esther?

MEMBER NEUWIRTH: Yes, so this one is widely in use but -- is it working now?

This one's widely in use. I don't think we had many questions in relation to it. But, I would say, you know, because it's widely used and tracked and has a long track record, I guess some of the questions that we had related

to the fact that lots of patients are excluded from this measure, so that seems like it -- some questions around applicability and use.

And then personally, I was wondering why this isn't a real PRO and I think that came up in some of our reviewer comments as well. Seems like there could -- this could very well be, you know, directly asked of patients rather than simply assessment because there might be some bias there.

Then it wasn't also clear how this really drove improvement efforts.

CO-CHAIR PARTRIDGE: Sherrie Loeb?

MEMBER LOEB: The pain one, I believe is also widely used. Pain is just a really tricky one to assess. You know, what's the worst pain for one person is minimal for someone else.

So, you know, we've got to track it. We've got to continue and home health is important to know what they're doing, what they're helping with, what the physician's prescribing at home is going to work.

But, whether we ever can truly find a tool to adequately assess it, you know, good luck. I hate the 1-10 scale. I hate the smiley faces, you know. Someday we're going to come up with --

MR. HITTLE: This David Hittle from the University of Colorado.

I should stress that we're not using the 1-10 scale. We're not using smiley faces. We're using how often does pain interfere with activities.

MEMBER LOEB: Right.

MR. HITTLE: And that's the scale that we're measuring folks on.

MEMBER LOEB: And that's so much better. I mean, you know, when patients come into hospitals and they're saying, you know, rate your pain on a 1-10 scale and it's happened to me, you know, at times. You know, if I'm going into an emergency room, my pain's usually 15.

So, I like the fact that, you know, if it's interfering, especially if you have home

health coming to your home, they're helping you with your function, with your ADLs and to get back to your --

You know, at least your baseline or hopefully better, so to help that it's not interfering and you're able to move on. So, I like that.

CO-CHAIR PARTRIDGE: Katherine?

MEMBER BEVANS: I just don't know how we're handling the multiple measures --

CO-CHAIR STILLE: We're discussing that.

MEMBER BEVANS: -- in a minute. Okay, I just want to suggest that I would like to talk a little bit about the appropriateness of other reporters for assessment of pain, in particular, so that we can consider that independently.

CO-CHAIR STILLE: Yes, we'll consider pain separately, I think.

MEMBER BEVANS: Okay, thank you.

CO-CHAIR STILLE: As long as everyone

feels good about that.

CO-CHAIR PARTRIDGE: It's my judgment as we've talked about these five that there are issues and, perhaps, at different levels in whether it's importance or feasibility or usability.

And probably we should consider voting one by one or if we could do it as a group but I wonder --

Okay, but I would be open to recommendations.

CO-CHAIR STILLE: Sure.

CO-CHAIR PARTRIDGE: Lisa and Chris?

CO-CHAIR STILLE: Lisa first.

MEMBER MORRISE: Let's just get on with it.

CO-CHAIR STILLE: Okay. I'd like to recommend that we talk about the first three as a group and then the last two med and pain separately.

I see some heads nodding.

CO-CHAIR PARTRIDGE: Okay. So,

Nadine, you want to work your magic?

MS. ALLEN: Voting on evidence, one yes, two no. Voting starts now.

For this measure, for the three grouped measures, can we do a hand vote? I'm experiencing some difficulties with the evidence slide.

So, yeses?

(A SHOW OF HANDS)

MS. ALLEN: Voting on performance gap, one high, two moderate, three low, four insufficient. Voting starts now.

Still waiting on a couple of votes.

All votes are in, 18 percent high, 76 percent moderate, six percent low, zero percent insufficient.

For evidence, the results were 17 yeses and zero no.

Voting on high priority, one high, two moderate, three low, four insufficient. Voting starts now.

So, all votes are in, 59 percent high,

41 percent moderate, zero percent low, zero percent insufficient.

Voting on reliability, one high, two moderate, three low, four insufficient. Voting starts now.

All votes are in, 35 percent high, 47 percent moderate, six percent low, 12 percent insufficient.

Voting on validity, one high, two moderate, three low, four insufficient. Voting starts now.

All votes are in, 29 percent high, 41 percent moderate, 12 percent, 18 percent insufficient.

Feasibility, one high, two moderate, three low, four insufficient. Voting starts now.

Still waiting on one vote.

All votes are in, 53 percent high, 41 percent moderate, six percent low, zero percent insufficient.

Voting on usability, one high, two moderate, three low, four insufficient

information. Voting starts now.

So, all votes are in, 24 percent high, 65 percent moderate, 12 percent low, zero percent insufficient information.

Voting on overall suitability for endorsement of measure 0167, 0174 and 0175. Voting starts now, one yes, two no.

All votes are in, 100 percent yes, zero percent no.

CO-CHAIR PARTRIDGE: But I do not want us to lose Peter's point. And the question is how we incorporate that in our comments somehow and perhaps CMS would like to come back to us to talk a little bit about the possibility of amending the exclusions of the denominator. That's one obvious way to look at it. There might be others.

But when this gets to CSAC, this issue will come up and both of us who are -- Ann and I are going to have to be able to respond.

So, I'm getting a nod from CMS, thank you.

MEMBER MORT: If I could just reinforce the importance of that comment.

And I looked through the OASIS, a 25-page assessment, and as physician, when I order home care services, I don't order everything. It's in some menu and I'm picking certain things that I want this particular patient to benefit from.

That might be some way in which you could identify what the goals of a home care episode are and monitor exactly who gets in the denominator.

So, to judge a home care facility on a criterion that I wasn't looking to improve just screws, you know, distorts it. And that would be a way, I think, Peter, to get at what you're concerned about.

If I'm ordering home care for nursing care to monitor congestive heart failure and ambulation is not an issue for this patient, the patient is where the patient is going to be. Then I think the home care services should be

judged on their performance on the medical issues, not on the PT issues. PT may not even come.

So, that's a way that I think you could address some of these really important issues.

CO-CHAIR STILLE: And when I agree, I mean from a child health standpoint, and these, you know, all should basically be extended down once we get them going, right, it's a relatively small number of kinds and adolescents that have ambulation as a piece of their home health care. It's usually a lot of different things, many things on this list, but, yes.

MEMBER THOMAS: Just quickly, no one is saying that it's not important, the measure improvement in home health services. That's not what I was getting at for sure.

It's very important for most people, but my only concern is that you put -- you don't want to create a system where you've got disincentives to treat the people who may not

improve but still might need that therapy in order to maintain or prevent deterioration of function.

Degenerative, you know, progressive kind of disorders or neurological diseases and things of that nature where they have every right to and Medicare is supposed to cover that care. And if there's a system in place that measures improvement alone without any recognition of that, there is a real risk that those people would be under served and that's -- it's not something we should do.

MS. MCMULLEN: Yes, I mean -- and I wanted to say something on behalf of CMS earlier and I'm with my Deputy Director for the Division of Chronic and Post-Acute Care.

We hear you and we agree with you and we are moving in the way where we are able to balance out the incentives so that we're not developing measures that are seen as being able to disincentivize care or take away from the actual goals of the resident or the patient or

the individual whoever that may be.

So, yes, we hear you and I have taken notes and we will be able to respond to that if need be at the CSAC or what not.

MEMBER THOMAS: So, we don't need to talk about this further every time this comes up. This applies to home health, SNF and outpatient. And whenever you've got that scenario, if you're measuring improvement, you've got to have some way to accommodate this settlement that CMS has signed on to.

Thanks.

MEMBER KAPLAN: I have one quick question. If we've endorsed these measures, is there -- we would still like to get some of the evidence that we've asked for on, for example, inter-class correlation coefficients, the means for the fit data for Rasch variables, et cetera, et cetera.

Is there -- how does that work?

MS. THEBERGE: The developers can provide it.

CO-CHAIR PARTRIDGE: I think it's always in the developer's interest where we've asked for it to provide it. It makes our case for recommendation stronger or weaker, but we hope strong, and particularly when we've got gray zones -- we don't have gray zones here, but where we do, it becomes very important.

Sherrie?

MEMBER LOEB: Just two quick comments.

I know that Liz who said as far as, you know, you recommend just a certain thing with home health, I think from the patient standpoint and from the nursing standpoint, when the home health nurses go out there, they may see things that the physician doesn't see in their quick visit.

And so, I mean I have had some -- a phenomenal home health nurse who came out and noticed things that maybe the internist didn't because they were only there for a short time. So they may pick up things they do need.

And, yes, you may have a CHF patient

who you want, you know, that taken care of, but who can't walk because of the CHF and so needs that kind of rehab, too.

And the other point is, if we know there's not going to be any improvement because of their chronic illness, that, unfortunately, hurts the patient in a lot of aspects. They a lot of times won't qualify for any more physical therapy. There's no improvement so we're not going to give them physical therapy. But what's going to happen without that physical therapy? Because they're going to go backwards.

So, the same thing can come into play with the home health, you know, visits. You know, there has to be some type of, you know, exclusion much like, you know, not everyone can get a beta blocker with cardiac disease, so there's a way to, you know, exclude it so you're not quote, dinged by accreditation.

So, in the future, somehow there has to be a little box so you're not hurt by not improving.

CO-CHAIR PARTRIDGE: All right. We've talked about the first three and we've talked briefly about 0176 and 0177.

Is there -- and Katherine wanted to talk a little bit about those further.

MS. RICHARD: So, do you want sort of the brief introduction of those other two measures?

MEMBER BEVANS: No, not necessary.

CO-CHAIR PARTRIDGE: No, we're comfortable with --

MS. RICHARD: Okay.

CO-CHAIR PARTRIDGE: Thanks anyway.

MEMBER BEVANS: Yes.

MS. RICHARD: No problem.

MEMBER BEVANS: It's really just, you know, as I mentioned, the point that the validity of other reports, other individuals reports of someone's pain has certainly been questioned.

And so, to the developers, have there been any validity tests done, evaluations of accuracy of the other reporters in comparison to

the individuals report of pain at the individual level?

CO-CHAIR STILLE: So, this is for 0177?

MEMBER BEVANS: Yes.

MS. RICHARD: This is Angela.

We've not done those particular studies and in the latest reviews, I haven't really seen a comparison of that.

MS. COOK: This is Keziah.

I think it may also be important to note that when the home health agency fills out the assessment item, the assumption is that they are drawing on multiple sources of information. They're drawing on their own observations of the patient, they're drawing on things that the patient has said and also potentially drawing on information from other family members or care givers.

So, they may directly ask the patient to what extent is pain interfering with their ability to do certain tasks, they're also

observing the patient.

So, if they notice the patient is really slower, hesitant to stand up and they're grimacing, that may, you know, be sort of evidence that pain is interfering with that patient's ability to do that particular task.

CO-CHAIR PARTRIDGE: Okay, further --

MS. RICHARD: Yes, that's very true. I'm sorry, that's right, multiple sources of information are used for the assessment of their function.

CO-CHAIR PARTRIDGE: Further comments or questions about either of these?

Lisa?

MEMBER MORRISE: In my experience, the ability of a patient to manage their own medications has a number of variables attached to it and I'm not seeing those necessarily here.

For example, when you start getting over a certain number of medications or a complicated schedule, it doesn't matter if you sit down and you do a chart and you do everything

possible. It's just you're not going to get adherence because of the complexity of the situation.

So, I'm looking at this measure and thinking, gosh, you know, if they have two medications then this probably works really well. But in the realistic scheme of things where I had one mom recently who posted a picture of something like 30 medications that her child had to take when he came home, no amount of health care intervention was going to ensure that this very adherent, compliant, trying to do the best for her child mom was going to actually achieve getting all of those medications in her child.

And I realize that the measure isn't looking at pediatrics, but it's analogous to the adult situation.

So, I'm questioning that maybe there needs to be some other variables that are looked at.

MS. RICHARD: it might be a caveat that we do have a medication cross list measure

that looks at education and it's also possible for the home health agency to request that a doctor does sort of have medication reconciliations.

MEMBER MORRISE: Can I just follow-up and say, medical reconciliation current consists mostly of going over a list. It does not help a patient in terms of how they actually take those medications.

I know when my adult child came home from the hospital a little over a year ago with ten medications to deliver by NJ tube, I practically cried because it was more than I could figure out. And all anybody ever did in terms of MedRec was to say you're taking this med and this med and this med and this med and that is not helpful.

CO-CHAIR STILLE: So, Lisa, that's a good point. I went into the validity section and kind of re-read it and it sounds as though that there's been a lot of validation by the fact that they use it a lot and that people find it useful

but not a whole lot of objective things, which would arguably be hard to do, although possible, you know, chart reviews, bottle, you know, bottle validation, stuff like that.

So, yes, that's a good point.

CO-CHAIR PARTRIDGE: Are we ready to vote?

Nadine, 0176 and then 0177.

MS. ALLEN: Evidence, one yes, two no. Voting starts now.

CO-CHAIR PARTRIDGE: This is a one at a time vote just for any confusion.

MS. ALLEN: This is for measures 0176.

All votes are in, 94 percent yes, six percent no for evidence.

Performance gap, one high, two moderate, three low, four insufficient. Voting starts now.

All votes are in, 31 percent high, 63 percent moderate, six percent low, zero percent insufficient.

Voting on high priority, one high, two

moderate, three low, four insufficient. Voting starts now.

All votes are in, 75 percent high, 25 percent moderate, zero percent low, zero percent insufficient.

Voting on reliability, one high, two moderate, three low, four insufficient. Voting starts now.

MEMBER KAPLAN: I just had the same comment I had for reliability that I had before which is, we don't have the facility level reliability inter-class correlation coefficients, et cetera, et cetera. So, as that issue becomes a uniform issue if these get approved and then we don't get those data because there's no incentive for the developer to give us those data.

MS. COOK: I'm sorry, this Keziah.

We did include our reliability statistics at facility. The beta binomial tests compared with in the between variation at the facility level. And we also reported test, retest, IPC.

MEMBER KAPLAN: No, it says at the measure level and the earlier --

MS. COOK: I mean, so what those are doing is they are calculating the measure at the facility level and then comparing those rates.

MEMBER KAPLAN: No, no, no, I understand that. But it's the measures that are being -- it's not the way you're going to use the score which is as a composite across the measures. That's what we're looking for.

MS. COOK: We're not using these as a composite. These are reported individually.

I'm sorry, I feel like we're being conflated with the previous group of measures maybe. Each of these measures are reported separately as rates both for public reporting and for quality improvement.

MS. ALLEN: Voting on reliability. Voting starts now.

Still waiting on one vote.

Please vote again.

All votes are in, 13 percent high, 56

percent moderate, zero percent low, 31 percent insufficient.

Voting on validity, one high, two moderate, three low, four insufficient. Voting starts now.

Still waiting on a vote.

All votes are in, six percent high, 56 percent moderate, six percent low, 31 percent insufficient.

Voting on feasibility, one high, two moderate, three low, four insufficient. Voting starts now.

All votes are in, 38 percent high, 56 percent moderate, six percent low, zero percent insufficient.

Voting on usability and use, one high, two moderate, three low, four insufficient information. Voting starts now.

All votes are in, 19 percent high, 56 percent moderate, 19 percent low, six percent insufficient information.

Voting on overall suitability for

endorsement of Measure 0176, Improvement in Management of All Medications, one yes, two no. Voting starts now.

Still waiting on a vote.

All votes are in, 100 percent yes, zero percent no.

CO-CHAIR STILLE: Okay, so let's have a brief discussion of anything else related to the pain measure. I know we've talked about that, there may not be anything left to discuss.

CO-CHAIR PARTRIDGE: Silence.

CO-CHAIR STILLE: I think we need a break, but I think we need to vote before a break.

CO-CHAIR PARTRIDGE: Nadine?

MS. ALLEN: Voting on Measure 0177, voting on evidence, one yes, two no. Voting starts now.

All votes are in, 100 percent yes, zero percent no.

Voting on performance gap, one high, two moderate, three low, four insufficient.

Voting starts now.

Still missing a vote.

All votes are in, 31 percent high, 63 percent moderate, zero percent low, six percent insufficient.

Voting on high priority, one high, two moderate, three low, four insufficient. Voting starts now.

All votes are in, 69 percent high, 31 perfect moderate, zero percent low, zero percent insufficient.

Voting on reliability, one high, two moderate, three low, four insufficient. Voting starts now.

All votes are in, 19 percent high, 56 percent moderate, 25 percent insufficient, zero percent low.

Voting on validity, one high, two moderate, three low, four insufficient. Voting starts now.

All votes are in, 13 percent high, 50 percent moderate, 13 percent low, 25 percent

insufficient.

Voting on feasibility, one high, two moderate, three low, four insufficient. Voting starts now.

Still waiting on a vote.

All votes are in, 50 percent high, 50 percent moderate, zero percent low, zero percent insufficient.

Voting on usability, one high, two moderate, three low, four insufficient information. Voting starts now.

Still waiting on a vote.

All votes are in, 31 percent high, 69 percent moderate, zero percent low, zero percent insufficient information.

Voting on overall suitability for endorsement of Measure 0177, Improvement in Pain Interfering with Activity, one yes, two no. Voting starts now.

All votes are in, 94 percent yes, six percent no.

CO-CHAIR PARTRIDGE: We are going to

take a break, a stretch. And congratulations for moving through a lot of measures in a difficult, fast way.

CO-CHAIR STILLE: Yes, great work. And then we'll start with -- do you want to start with the IRF ones or the AHCAs?

CO-CHAIR PARTRIDGE: I think -- I'm trying to remember what was yesterday's.

CO-CHAIR STILLE: AHCA measures, yes, okay. So, let's start with 2613 and 2612.

(Whereupon, the above-entitled matter went off the record at 10:44 a.m. and resumed at 10:59 a.m.)

CO-CHAIR PARTRIDGE: If I can -- I hate to do this everybody, but if I can pull us all back together. I'm going to turn - I'm going to just start talking and if nobody's listening then that's the way it is.

CO-CHAIR STILLE: Everyone, please sit down.

CO-CHAIR PARTRIDGE: All right, we are going to turn next --

We are now going to turn to our long-suffering colleagues who have been patiently waiting their turn while we rejiggered the agenda multiple times and proceed to discussion of 2613 and to -- I can't read from the distance -- 2612?

CO-CHAIR STILLE: 2612, yes.

CO-CHAIR PARTRIDGE: Okay. And our developers are here and please proceed.

MR. GIFFORD: So, my name is David Gifford, I'm a geriatrician. We're from the American Health Care Association and we have before you two different measures, 2613 and 2612.

This is Improvement in Self-Care and Improvement in Mobility. They are both based off of the CARE tool, self-care and the mobility components of the CARE tool that was developed and validated by CMS.

It calculates risk-adjusted change from admission to discharge for anyone admitted from a hospital to a SNF regardless of payer.

And I'm not sure how much you want me to go any more detail given the time and the pace

you guys are going. Better to answer questions from you all.

CO-CHAIR PARTRIDGE: Anything further you want to add right now or, if not, we will proceed to the -- we'll turn to our discussants.

CO-CHAIR STILLE: Sharon and Peter.

CO-CHAIR PARTRIDGE: Okay.

CO-CHAIR STILLE: Yes.

CO-CHAIR PARTRIDGE: All right, Peter?

MEMBER THOMAS: Well, I won't go -- belabor it at all, but the same issue about the Jimmo case, I think.

Although I do recognize there is risk-adjustment in this, am I not correct? I'm looking.

I found this to be fairly highly rated, well, the evidence in terms of this measure, I found it to be moderate.

I through the rationale speaks of independent living but the measures are pretty basic and I'm wondering whether -- I don't know enough about the CARE tool.

Did you select a subset of measures or is this all of the measures that would go to mobility under the CARE tool?

MR. GIFFORD: I think they're all the measures. We dropped one or two based on feedback from CMS that were measures that did not perform well in the PAC demonstration item. They are very similar to off the Barthell or any of the other items out there.

The mobility has 14 items, the self-care has eight items. They are aggregated together into an aggregate score that goes forward with that perform well.

MEMBER THOMAS: So, because there's a sizable portion of the SNF population that hopefully will go home and live independently, I have found that it'd be great if there were some measures that assessed the ability to cook and ascend stairs and things that I Just didn't see in these, more ADL and IDL kinds of activities that would enable someone to actually be mobile upon discharge from the SNF.

Other than that, I did find that there was a significant variation and room for improvement in terms of performance gap. There was no disparities data. I read that there's too low a sample size. Can you --

MR. GIFFORD: Overall, about 15 percent of the entire nursing home population in the country admitted to a SNF or of any sort of ethnicity group, they all tend to be concentrated in individuals.

There's a -- most of the studies that have looked at disparity work have shown that there is differences in disparity is to access to whether they go to a SNF or an IRF or an LTCH.

There's very little on the differences in quality in the SNF. Those that have looked at it, not in the ADL or mobility section have shown that there are disparities but they relate to the types of providers that are in -- that they tend to go to inner city, poor quality centers where both Caucasian and all ethnicities have equally bad outcomes.

So, if you looked at it at a population level, you'd definitely see differences where you don't.

We were following the strict guidance from NQF that if you are going to look at any ethnicity, you need to stratify that, you don't risk-adjust for it. And the sample size is so small in most nursing homes with only about ten or 15 percent being of any ethnicity that you would exclude over three-quarters of the SNFs in the country from the measure.

Certainly with renewed information out there, there may be a question about now risk-adjusting it with the changes in policy at NQF. But the guidance we got when we submitted this was to not to do that.

MEMBER THOMAS: Did you want to stop at reliability and validity and do that a little later or do you want to continue to go?

CO-CHAIR STILLE: Ideally, that would be great, yes.

MEMBER THOMAS: Okay. I mean the only

thing I'd say is that there is a, you know, what, 2.3 million Medicare admissions to SNFs in 2012 data, $229 billion spent on it. So, obviously, this is a very significant priority to try to assess measurement of mobility in SNFs.

CO-CHAIR PARTRIDGE: Any other comments on importance with respect to 2613? For example, gap.

MEMBER MONROE: Just one of the things that appealed to me about this measure, that it's regardless of care. And I think some of the other similar measures that we've seen have been payer-specific which I think creates another set of problems and encourages duplication of measure.

So, I appreciate that this is regardless of payer which makes it a more widely available use tool.

MEMBER THOMAS: And I'll also Just say that the American Occupational Therapy Association submitted some comments and I found them to be very helpful.

Where the description of this measure, you know, refers to that these patients are admitted for therapy in a SNF, that they made the point, well, they're really admitted for overall need and they received a significant amount therapy typically.

But that's not necessarily why they're admitted and they made that point. I just wanted to call that out.

Do you, you know, do you disagree with that?

MR. GIFFORD: Right now, 90 percent -- well, about 85 percent of all admissions to a SNF come from a hospital. Of those, 90 percent of them, whether they are in Medicare or not, end up in a rehab rug, which means they'll be getting PT, OT or speech therapy out there.

They often are there for additional information but one of the primary purposes for the skilled services is to improve their ADL function and/or mobility so that they can hopefully return home.

I think what you're touching on is you probably need a portfolio of measures to look at post-acute care, not just these measures including discharge to the community which we have a risk-adjusted measure we'll be bringing forward to NQF, it just doesn't go to this panel, it goes to a different group.

We're looking at satisfaction. We're looking at rehospitalization measure which just got approved by a different panel from NQF recently. But you almost need to look at those all in conjunction in any of the post-acute providers, not just SNF but IRF or LTCH or anyone else.

And I think your point is well taken that if you're looking at a portfolio for overall global care, you do, but we were limited to just looking at function at this point.

I think in your comments about the items in there, we wanted to use an assessment instrument that was in the public domain that was going to comply with the IMPACT Act that was

passed by Congress this last year and that CMS was going to use.

And since they had developed and modified it and we approached CMS, they indicated that they were leaning towards using the CARE tool. So, we elected to develop the measure around the CARE tool for that purpose on reason.

I think the additional ALDs or IADLs that you're talking about would make sense, but probably would be a separate measure because of the differences in the domain and the functional level of the individuals that were there. And so we were restricted to that domain. That would be a whole new sort of measurement development aspect.

The measure does look at some aspects to getting home which would be going up and down stairs. But some of the more IADLs like cooking and cleaning are not there but clearly you have to be able to do your ADLs to be able to get to your IADLs if you want to get there.

CO-CHAIR PARTRIDGE: I have a question

that really goes to the CARE tool.

Some of the other measures we've been discussing over the past two days include questions about the patient's memory and communication skills.

Are there questions related to those issues in the CARE tool?

MR. GIFFORD: So, we took the -- the CARE tool is a compilation of different standardized assessments. One of them is cognitive function.

We took the mobility and the self-care assessment tools and used those for this measure. We had long and lengthy discussion, we risk-adjust for cognitive status but we use currently the FIM score which is a validated score out of the MDS to do the risk-adjustment.

We elected to risk-adjust all for MDS items because they are standardized and reliable and have been tested. Deb Saliba who developed it can probably tell you a lot more about this, but we use that and so we risk-adjust for

dementia.

We had a panel of therapists that indicated that was a strong thing there and we include that in the risk-adjustment model.

CO-CHAIR PARTRIDGE: Good, thank you very much.

MR. GIFFORD: We're just not based off the CARE tool one.

CO-CHAIR PARTRIDGE: Right.

MR. GIFFORD: If CMS inserts that into the MDS, OASIS and IRF-PAI, we would switch over and use that at that time but come back to you with the modification of that, but, yes.

MEMBER THOMAS: Did that risk-adjustment model also include say not just Alzheimer's but something into brain injury or something along those lines?

MR. GIFFORD: It includes a FIM score which is overall sort of cognitive status, so we capture a lot of those. So, rather than do it by diagnosis by diagnosis we sort of did it more at that functional level with the FIM score.

CO-CHAIR PARTRIDGE: Thank you. Becky?

MEMBER BRADLEY: And I apologize, this may not be the right section to discuss this, but I couldn't find it in your submission and maybe I missed it, but since this measure is based off the post-acute care pilot that was done by CMS, how many skilled nursing facilities participated in that PAC that this research is based on?

MR. GIFFORD: This was not based off the PAC demo. We actually collected our own data. We trained about 500 therapists and had three corporations and did it across a number of SNFs and then validated it in about 600 SNFs across the country.

Since that time, we've got, what is it, Ellen, how many organizations? Forty-six therapy organizations representing over a thousand SNFs collecting this data across the country.

MEMBER BRADLEY: Okay, thank you.

MR. GIFFORD: But Ann can probably --

how many were in the PAC demo -- SNFs in the PAC demo? About 40 SNFs in the PAC demo, but that was for the validation of the instrument and tool. We used it to validate the measure here, used 600.

CO-CHAIR PARTRIDGE: Sharon, do you have any comments to add to those that Peter made?

MEMBER CROSS: No, he pretty much covered everything that I was going to comment about.

The only question that I had actually was related to feasibility, but I was trying to keep in line so, yes.

CO-CHAIR PARTRIDGE: Are people -- I'm sorry, Katherine?

MEMBER BEVANS: Yes, I have a question first of all about the mobility measure. Is that fair game at this point? Okay.

About the applicability of the tool in looking at, and this sort of I think, echos one of Peter's comments about the level of the items,

how sort of difficult they are in reviewing the Rasch coefficient.

It appears that the items are largely pretty good at measuring people on the upper end of the mobility continuum as compared to the kind of extreme low end and I'm curious to know your thoughts on whether you feel this mobility -- this combination of items is going to do a good job of sensitively discriminating people who have very low levels of mobility.

Is it going to be able to assess change among people who have low levels of mobility both in and at the end of their care?

MR. GIFFORD: I think the tool performs relatively well at those extremes. I'll defer to Ann Deutsch from CMS who developed the actual CARE tool to answer that in more detail.

But you see the items that's on page seven of the application, you can see the items there that rate on a 1-6 level. They held together and when we did a separate validation through Rasch, it looked and performed equally

well as in the PAC demo.

We did not use Rasch analysis to transform the scores on this. We used the raw scores based somewhat on the comments earlier and the fact that when we calculate this measure with the transformed sore on Rasch and we calculated it without a transformed score, the correlation was .98.

And so, to keep the measure simple for providers and members to understand and use it and be able to more calculate it, we elected not to use Rasch in the transforming of the data that was out there.

But, because again, as I think Sherrie's been pointing out over and over again, this is a measure of a provider level not individuals. And when you aggregate all this together, a lot of these questions that have been coming up actually become immaterial because it's aggregating at the provider level. This is a provider level measure as are the other measures.

MEMBER KAPLAN: Can I just follow that

up?

So, all of the reliability we still don't have the reliability at the facility level nor do we have the validity at the provider -- the variables here at the patient level for everything, correct?

MR. GIFFORD: So, we relied on the -- for the reliability, we did do what we could do with the availability of the data and use out there. We did a bootstrap analysis looking at it over and over again.

Yes, I agree, not perfect but it's -- we did something. No one else has done anything, we did something. You've got to give us some credit on something. Okay?

MEMBER KAPLAN: I'm all over it.

MR. GIFFORD: You passed everything else with nothing, so we did something.

MEMBER KAPLAN: I appreciate the effort and I'm not trying to smoke anybody. I just -- or set the bar too high.

But, you could do with the data you

have inter-class correlation coefficients, right?

MR. GIFFORD: Yes, we could probably try to go back and look at it.

We did look at the reliability across with other instruments because we did this in a number of the SNFs that were doing other instruments and it was very -- it worked with the Barthell and then the other two were sort of home grown instruments that they had done, worked relatively well.

As far as the validity, we validated the measure with the individual items. We didn't go back and validate them, we relied on the PAC demo where they had done the validation of that.

MEMBER KAPLAN: What we were looking for again and again is a thumb print that suggests that there is a reproducible score within facility that can then be used to compare between facility differences.

If there is kind of buckshot within the facility, that really weakens the ability to use these to compare facilities.

CO-CHAIR STILLE: Should we start voting? Becky, did you have -- okay.

CO-CHAIR PARTRIDGE: Are we doing it together? I'm sorry, I'm sorry.

CO-CHAIR STILLE: Together or separately?

CO-CHAIR PARTRIDGE: Together or separately?

MEMBER THOMAS: (OFF MIC)

CO-CHAIR PARTRIDGE: Yes.

CO-CHAIR STILLE: Yes. Right.

CO-CHAIR PARTRIDGE: Right, we're just talking about importance, two measures, is that what we're saying?

CO-CHAIR STILLE: I think there's a consensus for together.

CO-CHAIR PARTRIDGE: Okay.

CO-CHAIR STILLE: I don't see any different.

MS. ALLEN: Voting on Measure 2613 and Measure 2612, evidence, one yes, two no. Voting starts now.

All votes are in, 94 percent yes, six percent no.

Voting on performance gap, one high, two moderate, three low, four insufficient. Voting starts now.

Still waiting on a vote.

Please vote again.

All votes are in, 31 percent high, 56 percent moderate, six percent low, six percent insufficient.

Voting on high priority, one high, two moderate, three low, four insufficient. Voting starts now.

Still missing two votes.

All votes are in, 63 percent high, 31 percent moderate, six percent low, zero percent insufficient.

CO-CHAIR STILLE: Okay, now reliability and validity.

CO-CHAIR PARTRIDGE: Discussion on reliability and/or validity?

MEMBER THOMAS: There we go. So, the

measure is risk-adjusted and I found it to be qualified as kind of easy to administer and clear. Two data sets were used.

In terms of validity, the developers compared the mobility measure to a whole host of other ratings including the five-star rating, the Nursing Home Compare and then some specific measures like pressure ulcers, rehospitalization and the like and received or obtained kind of mixed results I would say, some good, some moderate in terms of correlation between those measures.

The one that I wanted to ask you about is was the rehospitalization measure because you expected a certain outcome and apparently it didn't develop that way. So, apparently there's -- was it a negative correlation between improved mobility and increased rehospitalization? Which really makes no sense.

MR. GIFFORD: Yes, probably that is reflecting that patient population and the level of improvement. We correlated with the NQF

endorsed rehospitalization measure and it was a different direction.

The correlation was small but it was different than we thought. We thought that if better improvement we'd see a lower rehospitalization, that's what the therapists and clinicians and the group thought. But we saw the opposite through our direction, I'm not exactly sure.

All the rest of the measures that were correlated the way they were with five-star either moderate -- mild or moderate.

The strongest correlation was with, as one would expect, was with discharge back to the community and it was very strongly and statistically correlated with being discharged back to, you know, which would make sense with both ADL and mobility that's out there.

I will add, we also looked at this measure whether it be a changed score and sort of to what you talked about before, Peter, or we used just a discharge score risk-adjusting for

admission and the measure correlated at .98 and essentially, that's the exact same measure.

MEMBER THOMAS: Okay.

MR. GIFFORD: And so, that's why we came with just this one measure for that. And what really drove it is that functional level, the admission level which is really highly correlated with, as we all know as clinicians, through your clinical coming in point.

And that rehospitalization piece is probably what was why we saw the differences, it's about the sort of the acuity that's out there, that if you're really sick, you're not going to see as much improvement and get as much therapy to do that.

It ended up being sort of a business, but kind of with anything, you can start making any explanations, so I'm not exactly sure why it was with the rehospitalization.

I will say that rehospitalization, you know, all the post-acute measures that we are working on and developing, how you address it in

the measures is really critical because it really throws everything out of whack.

MEMBER THOMAS: The exclusions, ventilator patients, persistent coma, quadriplegic, hospice, children accounted for about, you said in the data, 1.1 percent of the patients were excluded, is that right?

MR. GIFFORD: Yes, and we actually have national MDS data for all. But all four million admissions to the nursing home from CMS, about two and a half Medicare fee-for-service, the rest are non. And we looked at that and nationally, this would be about one to two percent for the exclusions would be excluded.

MEMBER THOMAS: So, just a thought --

MR. GIFFORD: And when we looked at it, whether it was concentrated, because a lot of those are sometimes concentrated in SNFs. There was only, I think, about five or six percent that had more than five percent of their admissions related to those exclusions.

MEMBER THOMAS: So, just a thought in

terms of the improvement standard not applying to SNFs. If you're already listing SNF-specific conditions like ventilator and coma, et cetera, quads, you might consider some conditions that are typically degenerative or aggressive in nature, some neuromuscular conditions where patients might need that maintenance therapy and not -- get dings for not improving, if you see what I'm saying.

MR. GIFFORD: Yes. No, and there was some discussion. We had a panel of therapists and clinicians reviewing and this was their list that they thought came up.

We had a lot of discussion about making sure -- the default was we wanted to include as many people in this measure even though many of the clinicians had the same concerns you did because we wanted to avoid the game-ability and the unintended consequence of starting to try to avoid it and moving people around.

There was a lot of discussion about

whether there should be some improvement in those individuals with degenerative diseases or not.

But we can certainly go back and look at that. It's a small population, it wouldn't change it that much.

MEMBER THOMAS: Other than that, I think the feasibility, over a thousand SNFs currently, you know, collecting this data using this measure, but, what are there, are there -- how many SNFs are there out there?

MR. GIFFORD: There is 15,326 SNFs, I believe, last count, though it fluctuates quarter to quarter.

And I think, as I said, the IMPACT Act requires the standardization of assessment tool and CMS has leaned towards, and included with the development of the CARE tool and our membership has fully endorsed the adoption of the CARE tool to be inserted into the MDS.

We would see this becoming very feasible once it gets incorporated in the MDS widespread. But in the meantime, we have a lot

of people and now that we developed this, a number of our members are switching over and starting to use this and the therapy companies are starting use it.

And we're starting to work on providing this measure should it get through NQF to managed care companies who are very interested in this.

MEMBER THOMAS: And finally, again, the AOTA submitted comments that talk about how mobility is one thing but is a real precursor for deciding whether a person is ready to go home and be independent. That involves things relating to sequencing and problem-solving and temporal appropriateness and memory and a whole host of other things that really aren't included in this measure.

You've already said that there are other measure sets that could add some of this?

MR. GIFFORD: Yes, I mean we're not saying this is the only measure for post-acute care. This is a measure to look at the

effectiveness of therapy on mobility and we're not claiming it any other way.

It does correlate -- I think if you're looking -- really the goal is how many people are going home and we do have that measure that's out there which is, I think, a better measure to capture all that.

And then I think there are other -- clearly there are other purposes for post-acute care and you would need other measures to look at that. I don't think you would smush them all into one measure.

CO-CHAIR PARTRIDGE: Becky?

MEMBER BRADLEY: Earlier on some of the other measures, we talked about the importance of inter-rater reliability and the training of the clinical staff that were collecting this information.

Can you describe a little bit how staff is trained and how you assess your inter-rater reliability among these measures?

MR. GIFFORD: So, one we relied on the

PAC demo and the test they did with reliability. In doing this, we developed a detailed training around the program, particularly where there were variations in the care focusing on where there were variations in the CARE tool from the current MDS scales or other scales they use.

We required them to all go through a webinar training. We taped the webinar training to go through. And then they had to take a post-test that was a detailed scenario-based testing and they then had to actually rate on the CARE tool for each of those.

I think there were, I think 12 clinical scenarios for mobility and 12 for self-care. And if you were going to be doing either setting, you could do both. And you had to get, I think, it was an 85 percent pass rate which means you had to score it exactly right.

And there was a lot of discussion amongst the therapists as to what would be right and oh, well, maybe I would rate it a little bit that way but we had a panel that did it and they

had to get it right and if they didn't they had to, you know, go back and get that test and do that.

And then once the therapists submitted their first assessment, it was reviewed by a coordinator for each of the companies to make sure that they were filling out the items correctly and that they had some basis for it. And they compared it to the other tool.

And so, before they did that, they passed that, they could not submit the data for this analysis.

MEMBER BRADLEY: So, is that a requirement ongoing for the organizations that they do that continued testing and training so that you continue to have inter-rater reliability going forward?

MR. GIFFORD: You know, it's an interesting question because we were just doing it for this. We envisioned that it would be incorporated into the training of the MDS which CMS does and everything else. We have provided

this training to CMS.

But what's more interesting, maybe Ellen can comment, is a number of the therapy companies and others have taken our training and using it and -- do you want to comment on how they're using that training to do that -- so and actually, it's sort of happening on its own. We weren't going to make it a requirement, but many are.

MS. STRUNK: Just to follow-up on what Dr. Gifford said, the industry and professions are very interested in trying to gather information and realizing that we're coming from a place where everybody has their own definitions of things and there's no ability to compare.

And so, we find ourselves in the situation we're in today with not being able to, you know, explain and justify the care.

So, there is a great interest in the industry out there to have something that is compatible across everyone. So, as Dr. Gifford said, we've had hundreds of therapists, it's a

three hour training.

It's free, it's available and we are just trying to highly encourage anyone who uses the tool to go through the training and go through it on a regular basis so that we can feel confident in the information.

MEMBER KAPLAN: I just want to correct myself because I looked harder at the validity evidence that you provided and it looks like, although we don't have the facility level reliability, if you use the patient level of reliability, variables that you've gotten correlations with other things that are purported to measure quality at this level are -- they are correlated in the range of a third to about half of the reliable variance.

So, I correct myself. I want to go on record to just make that observation.

CO-CHAIR PARTRIDGE: Are there further comments?

Are we ready to vote? Nadine?

MS. ALLEN: Voting on reliability, one

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

All votes are in, 43 percent high, 38 percent moderate -- sorry, 44 percent high, 38 percent moderate, six percent low, 13 percent insufficient.

Voting on validity, one high, two moderate, three low, four insufficient. Voting starts now.

We're still missing a vote.

All votes are in, 31 percent high, 63 percent moderate, zero percent low, six percent insufficient.

Voting on feasibility, one high, two moderate, three low, four insufficient. Voting starts now.

All votes are in, 25 percent high, 69 percent moderate, six percent low, zero percent insufficient.

Voting on usability and use, one high, two moderate, three low, four insufficient information. Voting starts now.

All votes are in, 31 percent high, 56 percent moderate, 13 percent low, zero percent insufficient information.

Voting on overall suitability for endorsement of Measures 2613, Care Improvement in Self-Care and 2612, Care Improvement in Mobility, voting starts now, one yes, two no.

All votes are in, 100 percent yes, zero percent no.

CO-CHAIR PARTRIDGE: And I want to thank the association for your patience, for your presentation, both your written materials and your presentation today.

We apologize again for keeping you for so long. But thank you for coming.

MR. GIFFORD: Thank you. The ends justify the means.

CO-CHAIR PARTRIDGE: We're going to give -- Sarah is going to give us a status report. How are we doing? Well, we cleared the first --

MS. SAMPSEL: So, we have five

measures left and they are split between inpatient rehab and long-term care. But it think it's all the same -- it's RTI and CMS.

So, we'll have them come up and we'll start and move through those and see how far we get.

We will be kind of keeping an eye out, you know, we fully recognize a few folks have to leave in the next hour or so and, you know, as was mentioned earlier, at the point that we lose quorum which is losing two more people, then we'll have the discussions and vote later. But, we'll see how these go.

CO-CHAIR PARTRIDGE: Okay, 2635 is at the intermediate rehab level and we're going to turn to Anne for her presentation or introduction.

MS. DEUTSCH: Great. So, high, I'm Anne RTI. Poonam is also with me from RTI and from CMS, we have Tara, whom you met yesterday and Stacy Mandl. And I think on the phone, Tracy Kline, are you there?

MS. KLINE: Yes, I am.

MS. DEUTSCH: Great, thanks.

So, for efficiency, I thought it would be helpful to kind of talk about what the item -- what the measures have in common and talk about CARE tool in general, clarify a few issues that have come up and then I'll get into talking about each of the different measures. So hopefully, that helps.

So, again, the items that we are talking about as part of these four measures for inpatient rehab facilities are care items. We have split the items up into self-care and mobility and the reason that we did that is actually consistent with what Dave Cella had mentioned earlier.

So, there was a lot of research done on a lot of functional assessment instruments in the past looking at adding motor-type items and cognitive items together as well as bowel and bladder.

And the research showed that those are

really different constructs or different concepts and they really need to be split up.

And that's especially important when you're looking at very heterogeneous patients across different types of inpatient rehab facilities and potentially long-term hospitals.

So, that's one of the reasons we ended up doing self-care and mobility as separate quality measures.

So, and, you know, cognitive, we absolutely believe is very, very important. We think that's actually should be measured in a different quality measure.

I also wanted to clarify that the way that the items are scored, for example, with walking, if somebody walked extremely well but they have a -- this individual, let's say, they have a cognitive problem, and therefore, requires supervision, the care score would reflect this cognitive problem.

So, even though we consider them kind of motor items, self-care and mobility, there is

a cognitive component to each of the items. If somebody has cognitive problems, they will not score independently on any of these items.

So, again, seven self-care items were selected. As mentioned earlier, additional items were tested as part of the post-acute care, post-acute payment reform demonstration and the items that didn't have good reliability or didn't really measure to the construct well were not included in this particular measure.

We have 15 mobility items to speak to. I think, Peter, you had a comment about range of ability being measured. So, we do have low functioning items, things like bed mobility and so we do go to a low end.

And then we also go to a high end by looking at things like whether the person can from a standing position, pick up an object from the floor. And that's really important because those are the kinds of things that individuals sometimes fall when they're doing those kinds of things after being hospitalized.

We also have things like car transfers which, now a days, is a daily activity whether the person is driving or not after they go back to the community or take a taxi, getting in and out of a car is important. So, we have some harder items also across.

So, we talked a little bit about testing yesterday. We presented in our materials the testing across all post-acute care settings as well as testing related to inpatient rehab facilities.

Again, the development including expert panels. We had patient representatives on each of our panels.

So, we have four measures. The first one is the mean risk-adjusted change in mobility score between admission and discharge for inpatient rehab facility and Medicare patients.

The second one is mean risk-adjusted change in self-care score between admission and discharge for inpatient rehab facility and Medicare patients.

The third one is the percentage of Medicare inpatient rehab facility patients who meet or exceed an expected discharge mobility score.

And the fourth one is the percentage of Medicare IRF patients who meet or exceed an expected discharge self-care score.

So, the items in the two self-care measures are the same. The items in the two mobility measures are the same. We're just aggregating the data differently at the measure level.

And the reason that we do that is based on previous research where I created kind of fake public reporting report cards. We went out to consumers and showed them mean change in function scores, self-care mobility. It was fictitious data that we made up.

And consumers really didn't understand what that meant. They understood measures that are reported as percentage much better.

And so, we felt it was important given

that there's different stakeholders that have interest in this data, that for consumers, the percent measures would really be more understandable. And for the industry inpatient rehab facilities in this case, looking at mean change in scores is actually what they are used to looking at. So, that's why we have the two sets of measures.

So, in terms of importance, I want to address, you know, this is the reason inpatient rehab facilities exist, improvement in function.

I think we mentioned in our application that there are differences in outcome in the literature that people have reported in terms of race/ethnicity, also length of stay has gone down in inpatient rehab facilities over time and function has actually gone down as length of stay.

So, it's really important to pay attention to the financial incentives that are lowering length of stay and could impact the functional outcomes of patients.

So, we don't really have a numerator denominator, so we -- similar to readmission measures that have been presented, we have an observed value that we create, an expected value that's a risk adjusted value and then we multiply the national average so you get a risk-adjusted value.

Our risk-adjustment model, we spent a lot of time, we've got a lot of feedback. We did a public comment and got tons of feedback from I think it was -- I don't remember how many people, but, you know, we have a document, 300 pages of public comment and everything that people suggested we test, we tested in our regression models to see if they were important risk-adjusters.

We considered both clinical, statistical literature when we put our models together. We were very thoughtful, I think, about our exclusion criteria.

So, for example, inpatient rehab facilities admit patients with locked-in syndrome

sometimes in tetraplegic complete. And those patients would not be expected to improve in walking and so, we have exclusion criteria when it is clinically appropriate these patients be excluded.

Our expert panel was very supportive of all of these exclusions because, as Peter mentioned, unintended consequence, we don't want people to have limited access because they're not going to improve on these particular items.

So, this is a clinician-rated instrument, not patient reported. CMS has historically provided training, I know that's come up. And as part of quality reporting programs has done auditing.

Let's see, I want to bring the work comments that were submitted, so I want to address those comments.

So, somebody wrote in and said it was data are based on a cross section study, that's not accurate. It's actually a prospective cohort study that we conducted. So, that's why we have

admission/discharge data.

We also link claims data so we could look at readmission rates and all that related to the data.

We did not use CART regression, so I'm not sure where that comment came from.

We used a generalized linear model with general estimation equations.

There was a comment that we only had 20 or 30 rain injury patients. In fact, we had 403, so I'm not sure where that comment is from.

Let's see, I guess Dave Cella made a comment earlier about the raw versus Rasch. So, I just wanted to address that because we chose similar to the previous measure, we chose to do the raw sum scores.

And that was based on research that I was involved in that looks at minute of the assistants and the raw scores, as Dave Cella suggested, actually were more correlated with minutes of assistance than the Rasch measures. So, happy to provide a reference if people would

like that.

But, basically, minutes of assistance is our gold standard. Minutes of assistance is ratio level data integral level so that is an appropriate gold standard I think.

I think that's it. So, did you want to add anything else?

MS. MANDL: All right, I've been asked, this is Stacy Mandl from CMS.

I've been asked to clarify on the submission of data for Medicare beneficiaries.

Currently, through regulatory requirement, through regulation, the data for the IRF-PAI to CMS is required on Medicare beneficiaries. And so, therefore, and the cohort for this particular measure is going to be Medicare for the IRF-PAI data that comes in.

Thank you.

MS. DEUTSCH: So, I'll pass it back to the chairs.

CO-CHAIR PARTRIDGE: All right. So, would we want to start with 2635 as our

discussion?

And it's Ann and Dawn.

MEMBER MONROE: I was concerned that it was Medicare-only. I think that's a limited look at the world and could result in a lot of duplication of effort on the part of the facilities, but that is what it is.

I want to ask you about, when I read this, I had trouble distinguishing what it was you were really trying to measure.

Was it how well the therapies or the work within the IRF had done by the time someone left or was it that the score when they left was a predictor of how they would do in the community?

And you said that in a couple of places, but if that's going to be the case, then I need to better understand why you think the five today is a predictor of success in the community and what you did to answer that question.

Because this one seems, in my mind,

that you're taking a longer view than some of the others we've seen which has changed from pre- to post-. Is that clear? Do you understand what I'm saying?

MS. DEUTSCH: So, we are looking at the change between admission and discharge.

MEMBER MONROE: Right, but --

MS. DEUTSCH: So, that's consistent with the other measures that have been presented.

MEMBER MONROE: But one of the things you say in here is that because that score at discharge, I thought, is predictive of success in the community.

MS. DEUTSCH: Right.

MEMBER MONROE: And so, I'm asking for the connection between the score at discharge and success in the community.

MS. DEUTSCH: Yes, so that was part of the general literature overview. And we did do analyses and maybe, Poonam, you can look this up specifically, but we looked at what the relationship between the discharge scores and

whether people went back to the community or not.

So, a patients who were --

MEMBER MONROE: Or how well they did in the community, is that different than going back?

MS. DEUTSCH: Well, so we reported like, I think in our supplemental material, the average scores of patients who went in the community versus those who didn't to show that relationship.

And that reference to doing well in the community is just part of our literature review. We didn't -- so that's just part of the evidence from the literature, not specific analyses.

Does that help?

MEMBER MONROE: Well, it answers my question, I guess. Because, to me, the outcome, when you talk about it as an outcome measure, you're really saying it's not patient outcome in terms of their success in their functioning in the community, it's the outcome is how they left.

Right? Which I would wonder is perhaps not an outcome measure but, I don't want to pursue this too.

MS. MCMULLEN: So, one of the many uses of this measure is to be able to assess an individual as they traverse the care continuum.

It's our idea of standardization, so you'll see that presented in the panel the last two days. We had two LTCH measures, we have these four IRF measures and under the IMPACT Act, you know, we have to standardize in all these settings and things of that nature.

These measures are developed in a way that from admission to discharge, we are able to see a change in functional independence in either self-care or mobility so that we, through these measures, have uniform outcomes so we can link all these outcomes.

So, if someone enters into an LTCH, they move into and IRF, they move into a SNF, they go into a home health agency and their service is in that setting, that we're able to

follow them and to assess the complexity within those changes.

So, simply at this very point in time, these measures were created for those linking purposes among other purposes. It was for the purpose of standardization.

So, you'll see that the one LTCH measure with the vent, but with the four RIF, they are created there so we have those uniform outcomes.

I don't know if that helps with some of the outcome and kind of conceptualizing the form measures and why they're developed, but it was for the sake of uniformity.

MEMBER DOWDING: I just have a query because this particular measure we're talking about, just for the benefit of the people around the room who may not have read it in detail, is that, with this particular measure, you're calculating an expected score on discharge on admission and then comparing their actual score on discharge to that expected score. And that's

different to the mean change in function or self care at discharge. That's very different in terms of conceptualization, and it depends a lot on the accuracy of the risk-adjusted expected models. It's not a normal risk adjustment. The risk adjustment is actually saying what do we expect this person to do on discharge, so that's quite a different conceptualization, if I've understood it right because I might not have understood it right. It's quite a complicated thing to get your head around.

MS. DEUTSCH: So there's four measures.

MEMBER DOWDING: Yes.

MS. DEUTSCH: So there's two self-care measures and two mobility measures. So for one self-care measure and one mobility measure, we're looking at change in function.

MEMBER DOWDING: Yes, because that's not this measure. This is the expected change --

MS. DEUTSCH: I was kind of describing all four together at first, and then you're

right. I mean, there's two that are change, and those are risk-adjusted values. And then we also, secondarily, have this percent of patient who meet or exceed a benchmark, which in this case is the risk-adjusted value. So there's one self-care measure for that and one mobility measure for that, yes.

MEMBER DOWDING: Yes. So why have we got four? Why don't we just have the two and change? Why have we got this very complicated expected risk-adjusted model, which, actually, for the discharge self care isn't actually that bad, but for the other ones the actual predicted values are pretty low. So I would be really, really unhappy if I was an IRF and I was being expected to reach a functional score which only predicted 30 percent of patient outcomes. So I don't understand why we've got four, why we don't have the two. And it's just, you know, why are we over-complicating things, in my view.

MS. DEUTSCH: So I guess, I mean, the data collection isn't different for the two

measures.

MEMBER DOWDING: No, I know, which, again, my reason is why are you calculating two different scores, which, essentially, you're using the same data? I just don't understand why we have four and not two. It's just a --

MS. DEUTSCH: Right. So as I mentioned, in other research that I've done we actually prevented change scores to consumers. I went out to some, you know, day programs for elderly, and the consumers that we presented, they did not understand what a change score was. They did understand quality measures that were reported as percentages, so I think it's really the issue that we have different audiences for these quality measures and so they have a different understanding and ability to understand some of these things, I think.

CO-CHAIR PARTRIDGE: Anne, it seems to me what you're talking about is we've got the same data, we can furnish either, we can present it to you either way, and it's really a question

for the ultimate user which way they want to go. And so I think I'm echoing Dawn's question: do we have to have separate measures, or can we have a single measure with the option to report?

MS. DEUTSCH: I wouldn't know how to fill out an NQF application and not make it complicated by doing it two ways, and so that's why we presented it as two different quality measures.

CO-CHAIR PARTRIDGE: Helen and Sarah?

DR. BURSTIN: It's a good question. I mean, it probably is a complexity of our form that makes that hard to do. We do sometimes have measures that are submitted with two rates in the same measure. I think, just given the complexity that I think was just pointed out, that's probably hard to do. I mean, one question might be should they at least be paired so that you at least have an option of when you see one you see the other, which might be one way of getting around the idea of looking --

MEMBER DOWDING: And that also goes to

my anxiety about what then these are used for because if you have two different measures, they're using the same data but just analyzing it slightly differently, which one of the IRF facilities are actually going to be held accountable to and which one their pay going to be accountable to? It can't be both because you're using the same information. And, again, I would be very worried about the expected change one, them being held accountable to that one, because of the variation in the predictive models that you're using.

MS. MANDL: This is Stacy from CMS. I'd like to jump in on the policy question. These measures aren't used for pay for performance. They're not, we're not seeking endorsement for measures for use for pay for performance. That's a whole other ball of wax. It's the same data elements used to be able to calculate the two variations of the same measure. Consumer feedback, which moves beyond what Anne has suggested, is it's an exponential request

that the consumers be able to use this data. The facilities need the data, as well, for other purposes. They're very separate but both very meaningful.

But I just wanted to just make sure that the data coming in and the completeness of the data coming in and the accuracy of the data coming in is of the same data. So I just wanted to clarify. These are penalty for failure to report programs. Thanks.

MEMBER BIERNER: So I don't have a concern with it. I think it will be very useful for us, as an institution, to report to our consumers, the clients that we seek to attract, to present it in the percentage format or the benchmark kind of format for public purposes and then the other format for internal or for reporting purposes.

So I like this score overall, but my question is compare for me what you see as the differences between this and what we've heard earlier from the uniform data system measure,

which is using this Rasch format. Compare to me how you think yours is better or in what way it's different from that.

CO-CHAIR PARTRIDGE: We aren't in that discussion, but, if you want to answer the question, go ahead.

DR. BURSTIN: We typically look at comparison to other measures after something has been approved, so we probably don't need to do that right now.

MEMBER BIERNER: Well, then tell me about the six-point scale that you use.

MS. DEUTSCH: Yes, sure, sure. I'll talk about the instrument, the items. So six-level scale. Six is independent, five is setup or clean-up, level four is supervision/minimal assistance, level three is moderate assistance, level two is substantial assistance, and level one is dependent. Basically, as part of the post-care payment reform demonstration and instrument development, we had a lot of expert panels. We reviewed all of the existing

instruments: the OASIS, the FIM, IRF-PAI, you know, MDS. We asked for input. We also looked at literature, you know, what fit within what construct, what was reasonable in terms of scoring.

So for example, we got a lot of feedback that supervision takes a lot of time with patients, especially at the time of admission, the cognitive scores. So supervision is actually coded level four on the CARE tool, which is different than the FIM. And that was the rationale.

Also, you know, the FIM allows up to 25 percent effort from the patient for level one. But in talking to clinicians, they felt that having a dependent score that was truly somebody dependent was really important. So then if somebody came in totally dependent, perhaps, you know, minimally conscious, and then they made some improvement, there would be documentation on the CARE tool. They would go from a one to a two.

And so, you know, again, this goes back to what Peter said. Somebody who comes in very low level, are you going to see improvement? And in this case, that was the feedback that we heard from clinicians. They wanted to have that. So does that help?

MEMBER BIERNER: Yes.

MEMBER BRADLEY: And maybe I misunderstood the representative from CMS that addressed the payment. Did I understand you to say that, for payment for IRF, they would continue to use the IRF-PAI? Would the uniform -- with the FIM embedded in that and going forward. So is that what you said? I'm sorry. I didn't understand that.

MS. MANDL: I'm not sure I'm understanding.

MEMBER BRADLEY: So right now, for payment purposes, there is a tool, the IRF-PAI, inpatient rehabilitation hospitals are required to use, and it has a scale already embedded in that that drives our payment system.

MS. MANDL: No. What I was referring to is the question about Medicare only. Our data source is the IRF-PAI, and it's submitted, the regulation dates way back for multiple purposes, including care planning, as well as payment. But that's the regulatory requirement, and that's what drove -- when responding to the questions on the form for NQF, what's the population, it's Medicare. So that's all I was referring to.

MEMBER BRADLEY: So I guess my question is is that scale going to continue to be used? Because here this, I guess, not so much competing measures but inter-rater reliability. So if they're using a seven-point scale for the payment on the IRF-PAI and then a six-point scale of the same types of items and collecting different measures, it seems like it would be very confusing and the possibility of clinicians getting confused about do I use a six-point scale here or a seven point because the measures are so similar but different.

MS. MCMULLEN: Right. So we can't

speak to what will happen with IRF. As it stands today, we have the scales that we use. We've developed these measures, the CARE measures, to see, and even out of the Post-Acute Care Payment Reform Demonstration and the development of the CARE tool, the idea was to see if standardization could happen between multiple different types of assessment items within multiple different types of assessment instruments. And we have, CMS has been moving in the way with RTI as our contractor to be able to develop measures to see if standardization is possible. So that's where we are today is we're presenting these measures for consideration of endorsement, and these are the first set of standardized measures that we've been developing.

MEMBER BIERNER: Can I just say that the six is actually more intuitive for most of us who've worked in rehab and for a lot of therapists and others than the seven have been, so, actually, the terminology is more conventional. So I like the six.

CO-CHAIR PARTRIDGE: All right. I see Sherrie's card and I see Peter's hand up, so Sherrie?

MEMBER KAPLAN: This is, you know, the same old, same old. So on table one, these are average Cronbach's alphas at the patient level by facility. They're not the ICCs that we've been talking about.

MS. DEUTSCH: Correct. So our data are from the Post-Acute Care Payment Reform Demonstration, and we only had data for either six or nine months from the facilities, so we weren't able to compare data over time. The measure time frame is 12 months.

MEMBER KAPLAN: But you said something about generalized estimation equations, and then I was thinking, well, where are the splines? You know, where are the nice little facility-level error bars on the splines?

MS. DEUTSCH: We can get those for you.

MEMBER THOMAS: I want to go back to

the two different measures that you are using for both mobility and self care. So it seems to me, and I just want to get your point of view, it seems to me that it's probably, it seems to me to be more accurate to assess the change in function between admission and discharge, assuming that some gaming, I suppose, could take place, etcetera, than to come up with an expected functional level and then see how you achieve that, whether you exceeded it or didn't exceed it or hit it. So it strikes me as being less reliable to introduce that kind of subjective, somewhat subjective, kind of expectation of how well a person is going to do, and it's got all to do, I'm sure, with all kinds of comorbidities and all kinds of things.

And I guess my concern is what's to prevent providers from expecting low and exceeding high and looking really good in terms of how well they do?

MS. DEUTSCH: So to address your first comment, so certainly looking at change. We've

got much more granular data. We're taking advantage of every piece of data that we have. So you're absolutely right, there is more detail.

I think the percent measures, it's just to make it more understandable to consumers, again. So I can't speak to reliability. We didn't have enough data to be able to test it. And, certainly, I think that it's a great question and we should do that. But in terms of the expected value, I mean, the self care, I think, is like 85 covariance. And we look at comorbidities. We look at, you know, the clinical condition. We included interactions in our models. We tried to adjust for as much as we could. Our mobility, I think, has 85 covariance and our self care has, like, 74. So, I mean, we did a lot of work on the risk adjustment to do as good a job as we can. Does that address it, Peter?

MEMBER THOMAS: It does. Thank you.

CO-CHAIR PARTRIDGE: Are we ready to vote? Yes? The staff's suggestion is that we

take a look at two by two, 2635 and 2633, both of which are the met or exceeded with respect to self care.

MEMBER DOWDING: Can I just clarify? Are we going to discuss reliability and validity of these particular measures after --

CO-CHAIR PARTRIDGE: Yes, importance first. Is that acceptable? Okay. The two that relate to -- no, no, I'm sorry. Met or exceeded and the other self care one.

MEMBER BIERNER: Self care, self care.

CO-CHAIR PARTRIDGE: Self care, self care, yes. I'm having trouble reading my own notes. Okay. Nadine?

MS. ALLEN: Voting on measure 2635 and 2633, discharge self-care score and change in self-care score evidence: one yes, two no. Voting starts now. All votes are in: 94 percent yes, 6 percent no.

Voting on performance gap: one high, two moderate, three low, four insufficient. Voting starts now. All votes are in: 13 percent

high, 75 percent moderate, 6 percent low, 6 percent insufficient.

Voting on high priority: one high, two moderate, three low, four insufficient. Voting starts now. All votes are in: 56 percent high, 38 percent moderate, 6 percent low, zero percent insufficient.

CO-CHAIR PARTRIDGE: Scientific acceptability. Either. Reliability and validity, Dawn?

MEMBER DOWDING: Okay. Well, this is specifically with 2635, which is the one that I was asked to review in detail, which is the expected versus actual discharge, self-care discharge scores. And I just, I mean, I read through it a couple of times and you've got really good data on the reliability and validity of the CARE scale. I mean, we've demonstrated, I think, fairly comprehensively that it's a reliable and valid way of connecting data, but that's not the measure we're being asked to endorse. We're being asked to endorse your

calculation of a score related to somebody's function at discharge compared to what their expected score would be based on this risk-adjusted predictive model, and I couldn't see any data on the reliability and validity of that measure.

MS. DEUTSCH: So let's see. We sent some supplemental information that had the relationship between discharge scores and discharge to community. I'm sorry. Oh, and also the relationship between the CARE scores and length of stay, so that kind of speaks to some validity issues. At the scale level, you know, we presented several scale-level analyses. So, like, the discharge score is basically what the items as a group together and whether people meet that benchmark or not.

So we provided reliability. I called it scale level because it's the group of items. So does that help?

MEMBER DOWDING: Well, I mean, I don't have a problem with the score or the status.

It's this whole distinction between the scale that you're using to develop the measure and the measure itself, and the measure itself is this expected versus actual function at discharge. And that's the measure we're being asked to endorse. So we're not being asked to endorse how good the CARE scale is. We're being asked to endorse your ability to measure this at discharge.

And I didn't see the additional data, so I apologize. Can we bring it up so we can have a quick look to see how the -- because that's predicted validity, but there's still no data on the reliability of that measure. Does that make sense?

MS. DEUTSCH: So are you asking for facility-level reliability?

MEMBER DOWDING: Yes.

MS. DEUTSCH: Yes. So we didn't -- yes, this is similar to what Sherrie asked. I mean, we don't have multiple years of data, so we're not able to do that. So, I mean, I think

NQF provides guidance that the measure can't be high, but it certainly can be moderate if you provide item level and scale level --

MEMBER DOWDING: And I think that the care-level data is exceptional. It was more a query than anything else because, I mean, conceptually, I'm just having trouble getting my head around this expected versus actual. And for this particular measure, the actual predictive model, I think, is 0.53 prediction, so it's not too bad. I would be concerned about the other one.

MS. DEUTSCH: Great. Believe me, I'd love to have more data.

MEMBER DOWDING: Yes, I'm sure you would. I'm sure you would. I mean, it would help if we could see the predictive stuff.

CO-CHAIR PARTRIDGE: We're all taking a deep breath up here because I think this issue has come up 16 times over the past two days, and we're all extremely uncomfortable with trying to assess a measure at the level of a specific

provider entity without having that data.

MEMBER DOWDING: Yes, I mean, I sort of feel quite uncomfortable about being asked to endorse a measure for which we have no reliability data. I mean, it's just, you know, the scale measure is one thing, but it's going to be used at the facility level as a measure and being fed back to patients, and we don't have the data to assess it. It's not, it's just not there.

CO-CHAIR PARTRIDGE: And in some situations, we have had the capacity to go back and get it. In this particular situation, I think it's very clear we don't.

MEMBER KAPLAN: Can I weigh in here, though? Because --

CO-CHAIR PARTRIDGE: Sherrie?

MEMBER KAPLAN: -- they do have some data. You've obviously got generalized estimation equation data that have splines with error bars around each facility's performance that, actually, we could use to see how good we

are at, you know, discriminating versus errors in the distribution of facilities. So if you could give us those data, that would actually help a lot.

MS. DEUTSCH: Sure. And, I mean, we did present facility-level data, a mean, median, range, from the facilities we had.

DR. BURSTIN: A brief comment. I mean, I think sometimes what we see, as measures get out into use, we start getting a lot of this information. And so, you know, as we continue to explore what are requirements are for these issues, it's not simple, by any means. You know, one question might be, for example, by the annual update or something like that, can we ask that some of these additional data be brought back. You are a standing committee. You will still have standing in a year. There may just be ways for us to fully recognize there's only so much they can when the measures are not yet in play.

MEMBER KAPLAN: But aren't you guys exploring at NQF tiering approvals, too, levels

of, like, you know, phase one, you know, tiering for these purposes or those, generating data versus accountability?

DR. BURSTIN: Not yet. Right. And that's, I think, what we're finding to be the rub these days is that, you know, with the increasing influence of pay for performance, the anxiety about pay for performance, there is higher expectation of wanting to see some of those data. We get that, and that's what we're exploring.

But, again, there are opportunities, with you being a standing committee, to potentially get that information as they gather it and bring it back to you for additional re-review at a later date.

MEMBER BRADLEY: Could I just ask then, so it sounds like you're hoping to gather more data. How many IRFs do you have using this tool right now and how will you collect that data so that you have enough data to bring back?

MS. MCMULLEN: So we can't speak to future direction of measurement use at this point

because, as you know, I mean, with the IMPACT Act or whatnot, we're in a tough, difficult time line, and CMS is basically in the planning stages of figuring out how to specifically meet that mandate. At the current time, the data that we do have is from the CARE tool, the PAC PRD demonstration. It is our full intent in the future to be able to collect data on this so we can come to the table with measure-level data to be able to present a more accurate case for the outcomes and how appropriate the outcomes are for this specific area and domain.

But there's a lot, there's a lot up in the air right now because the IMPACT Act is not only our only mandate that we're faced with right now. We were given about two mandates within a month and a half. So to the future direction at this time, we can't speak to, but when CMS knows a future direction they will absolutely make that known.

MEMBER BRADLEY: Okay. So I guess then you really don't have the capability of

collecting a lot of the ongoing data at this point in time.

MS. MCMULLEN: I don't know if it's about the capability. I think the capability is always there. It's just the plans for CMS and what CMS intends to do in terms of collecting that data, and I -- definitely above my paygrade -- can't speak to that. Sorry.

CO-CHAIR PARTRIDGE: The question before us, again, and it's common to both measures, is we don't have facility-level data. They can supply some data which might make Sherrie Kaplan happy and, thereby, perhaps some other member so of the Committee happy. And we have faced this issue before. So I think, unless people object, we're ready to proceed on. Again, it would be both. And Nadine? We believe it works for both because it's the common issue. The facility issue is common, yes.

MS. ALLEN: Voting on reliability: one high, two moderate, three low, four insufficient. Voting starts now. All votes are in: zero

percent high, 47 percent moderate, 13 percent low, 40 percent insufficient.

Voting on validity: one high, two moderate, three low, four insufficient. Voting starts now. All votes are in: 7 percent high, 47 percent moderate, 7 percent low, 40 percent insufficient.

Voting on feasibility: one high, two moderate, three low, four insufficient. Voting starts now. All votes are in: 27 percent high, 53 percent moderate, 20 percent low, zero percent insufficient.

Voting on usability and use: one high, two moderate, three low, four insufficient information. Voting starts now. All votes are in: 20 percent high, 47 percent moderate, 20 percent low, 13 percent insufficient.

Voting on overall suitability for endorsement of Measure 2635, discharge self-care score for medical rehabilitation patients, and Measure 2633, change in self-care score for medical rehabilitation patients: one yes, two no.

Voting starts now. Still missing a vote. All votes are in: 67 percent yes, 33 percent no.

DR. BURSTIN: Any of the developers who have to stay around for this, please feel free to partake.

MS. SAMPSEL: Operator, can you open the line for public comment?

OPERATOR: All right. If you'd like to make a public comment, please press star and then a number one. No, no public comments at this time.

MS. SAMPSEL: Okay. Thank you very much. We're taking a break for lunch.

(Whereupon, the above-referred to matter went off the record at 12:23 p.m. and went back on the record at 12:55 p.m.)

CO-CHAIR STILLE: Welcome back. Let's get started.

CO-CHAIR PARTRIDGE: We are going on to voting. I think she's going to take root here.

CO-CHAIR STILLE: Sure. Yes, I think,

you know, we need to have some just sort of discussion about stuff, but then we're voting. No, I don't think we need any more from you, I don't think.

CO-CHAIR PARTRIDGE: But don't go away.

CO-CHAIR STILLE: But don't go away. We appreciate your presence.

CO-CHAIR PARTRIDGE: Yes, that's our third.

CO-CHAIR STILLE: Yes.

CO-CHAIR PARTRIDGE: It's the last one?

CO-CHAIR STILLE: It is. It is the last one. Okay, all right. The first one we're going to consider is -- do we have a quorum? We have 14. They're just somewhere in the room eating. Okay.

The first measure we'll consider is 2634, inpatient rehab facility functional outcome measure change in mobility score for medical rehab patients. And I and someone who is sick are the primary discussions, so I'll talk about

that.

Essentially, it looks at the change in mobility score between admission and discharge, does risk adjustment, and that's for IRF Medicare patients. So all those clauses and caveats in there. And let's see. In general, I'll just sort of summarize some of the comments and add a couple of things myself. The rationale for having a measure of this is important medical rehab for almost everyone. Enhancing mobility is a primary goal, so it seems to be supported.

There is a good score distribution and definitely room for improvement, given the numbers that are there. And there are some disparities that are also mentioned with race, ethnicity, insurance type, and region of the country. So that is good.

Other things. Well, there's a lot of validity and reliability things that we could talk about, but I think, in terms of the face, you know, it's Medicare patients only, so that's a little bit of a limitation in terms of the

group. It's IRF only, but it is what that is. And so, in terms of importance, I think it's got, you know, pretty good stuff on the surface.

So other thoughts and comments about importance?

CO-CHAIR PARTRIDGE: Again, this is the same pair of measures as we discussed just before our lunch break. The companion is 2632.

CO-CHAIR STILLE: Thirty-six?

CO-CHAIR PARTRIDGE: I'm sorry. Thirty-four and thirty-six. Only in this case, we're talking about mobility, rather than self care.

CO-CHAIR STILLE: Right. So 36, which I didn't review in detail, is the difference from expected, I guess, right?

CO-CHAIR PARTRIDGE: Okay. Any comments or questions? Any comments or questions? Peter?

MEMBER THOMAS: Okay. So can we just assume that all of the prior discussion we had on the self care that applied, you know, just kind of incorporate by reference, so to speak?

CO-CHAIR STILLE: Sure.

MEMBER THOMAS: I'm not sure that fully does justice to airing those issues, but I had the same basic concerns that I had in the first one with this one and the same questions. They weren't all concerns. They were just.

CO-CHAIR PARTRIDGE: Becky?

MEMBER BRADLEY: I just wanted a little bit of clarification. Since it's Medicare only, were the disparities you found related to insurance payments? I wasn't sure how -- was it within the insurance, like managed Medicare versus Medicare, or could you just address the disparity?

MS. DEUTSCH: All right. Can you hear me now. So that was literature that we cited, so Tim Reistetter at University of Texas Medical Branch did that study and he looked at, basically, public versus private. So it wasn't from our data. That was from the literature.

MEMBER BRADLEY: And just help me understand because some of the measures aren't

limited to Medicare patients with the CARE tool and some are. Can you explain why this one is and some of the others that were presented earlier are not?

MS. DEUTSCH: So the IRF ones are limited to Medicare only, and, again, Stacy provided the rationale. The Long-Term Care Hospital Quality Reporting Program was established as an all-payer program, and so, obviously, when we're able to get all-payer data, we want the measure to be all-payer. So anything -- no, that's it.

CO-CHAIR PARTRIDGE: Are there any other questions or comments before we vote on importance?

CO-CHAIR STILLE: Peter?

MEMBER THOMAS: Forgive me. I just wanted to say, in case people weren't aware of it, that you talk about these measures as being quality measures or pay-for-performance measures, but I think there's an assumption by many people around the room, probably rightly so, that

today's quality measure will eventually become or is likely to become some kind of a pay-for-performance measure. But in IRFs, there's currently requirements for pay for performance. There's a two-percent reduction in payments if you don't submit certain quality data, and I don't know how these specific measures factor into that, but I was trying to figure out what that connection might be.

MS. MANDL: So the Inpatient Rehabilitation Quality Reporting Program was established with Affordable Care Act, Section 3004. It is a penalty for failure to report. It's not a pay-for-performance program. So I just wanted to clarify that. So does that --

MEMBER THOMAS: It's helpful. That's helpful. The larger comment I made, I think, still stands, but thanks for clarifying that.

MS. MANDL: Sure.

CO-CHAIR PARTRIDGE: Are there further questions? Are we ready to vote?

MS. SAMPSEL: For a quorum right now,

we need 13 of 17, so we had two members that are not participating today. We're still okay. The doors are now locked.

MEMBER BRADLEY: Are we voting on these as a pair, like we did --

CO-CHAIR PARTRIDGE: We are voting on these as a pair.

MS. ALLEN: Voting on evidence for Measure 2634, change in mobility score, and Measure 2636, discharge mobility score evidence: one yes, two no. Voting starts now. All votes are in: 100 percent yes, zero percent no.

Voting on performance gap: one high, two moderate, three low, four insufficient. Voting starts now. All votes are in: 23 percent high, 62 percent moderate, 15 percent low, zero percent insufficient.

Voting on high priority: one high, two moderate, three low, four insufficient. Voting starts now. All votes are in: 54 percent high, 46 percent moderate, zero percent low, zero percent insufficient.

CO-CHAIR STILLE: Okay. So now we'll talk about reliability and validity. I think there's a little bit more to talk about in that way, although some of the discussion from the last thing is important.

Generally, you know, for 2634, inter-rater reliability was calculated a couple of ways. Item and scale reliability were done and were good. It was tested in a variety of venues. There was content validity that was tested against several other instruments for most of the items. A few didn't have a counterpart.

Then I just wanted to kind of go over some of the comments from the rest of the Committee, too, as well as an external comment, which was important. I think Dawn's concern about 2635 extends to 2636, given the expected versus actual score calculation. We don't have good data about that, I don't think. And there was no detail on exactly how that was developed.

There was a critique from UDSMR that was concerned with test-re-test reliability.

Maybe the developers can talk about that because I know you talked about some of their other concerns for the other one. And maybe I'll stop there and then talk about the validity concern.

Are you ready to talk about that quickly? You need a minute? Okay, I'll talk about -- oh.

MS. DEUTSCH: Okay, great. So we presented inter-rater reliability, so I guess I'll just generally speak to that overall. So the first inter-rater reliability is kind of this traditional where we had two therapists or two nurses go into a patient's room. They both did an assessment. So let's say Poonam and I worked at the same facility. We would go in and do the assessment. The instructions were people could not talk, but then we'd both independently score, and that was compared. Laura spent the evening yesterday and did a lot of the analyses, and we did both weighted and unweighted kappas.

Overall, that was good. We definitely, in the PAC demo, had some items that

didn't test as well, and those are not being proposed. So, you know, I just want to be sure you understand that we tested like a whole, a very long instrument, data tool, but not everything is being moved forward. Because perhaps Poonam and I were both trained at the same facility, you know, whether Tara and Stacy at another facility might be scoring the same was important. That's why we did the standardized patient videos, reliability, so that we were able to test if the four of us all came to the same agreement, and those of us who were PTs, nurses, you know, did we agree with each other? So that's why we did different types of reliability, so I feel like we've done a fair bit of reliability and our results were very comparable to anything that's out there in the literature with the existing.

CO-CHAIR STILLE: The only thing that I didn't see that was, the only thing I saw in the critiques that I didn't see in your thing was test-retest reliability. But, yes, I agree there

was lots of other reliability things that you guys --

MS. DEUTSCH: Yes. So with function, a patient can change. Like somebody with arthritis, they can be very limited in the morning and very independent in the afternoon. So function changes, so you can't really do test-retest reliability when somebody is changing. So I think that's not something that's generally done with function data. It's just not possible.

CO-CHAIR STILLE: Yes. And, yes, Peter and then Dawn.

MEMBER THOMAS: Dawn first.

CO-CHAIR STILLE: Well, Dawn first and then Peter. Okay.

MEMBER DOWDING: Okay. I'm just going to sound like a stuck record again. It's, again, just to highlight my concern that the reliability and validity data we have is at the level of the care scale and not the measure we're being asked to endorse, which are the two, the expected versus actual and the mean change. And we don't,

you know, we need the data on that, I think. Well, I need that data on that to be happy to endorse it. And, again, just to reiterate those concerns.

MS. DEUTSCH: Can I respond to that? So just maybe I can ask the NQF staff. My understanding is that the rating, if you don't have facility-level data but you have item reliability data, that it can be moderate. It just can't be high. So this is very acceptable, and a moderate would be the appropriate. Is that --

MS. SAMPSEL: That is correct. So that's consistent. This is an outcome measure. These are both outcome measures, not patient-reported outcome measures, meaning that when you go through the algorithm your choices would be moderate and low.

CO-CHAIR STILLE: Okay, good. And then Peter. Right, sorry.

MEMBER THOMAS: I'm sorry to do this. But can we just return to this expected

functional level in terms of mobility and just walk me through a little bit more, just give me a little more comfort that there's a system in place where you really are not going to find a lot of gaming going on or that you accommodate -- how do you determine what that expected functional level is? How does that happen?

MS. DEUTSCH: So the expected score is calculated based on our regression model. So, for example, you take the characteristics of the patient and then, you know, if the patient is in a certain age category, then you apply the regression coefficient that we reported on our risk adjustment model and, basically, you know, add up the scores that you get and the intercept and based on whether people have comorbidities or not, and the expected score is calculated based on summing the regression coefficients based on that person's characteristics.

MEMBER THOMAS: And, like, what was the sample for that, for those coefficients to be created?

MS. DEUTSCH: Right. So we had, after the exclusion criteria, we had what? About 4,776.

MEMBER THOMAS: Okay. And that's all considered valid and reliable in terms of expounding upon those, the experience of that sample group?

MS. DEUTSCH: Yes. I mean, we ran the analyses and we feel the estimates are pretty stable. I mean, we would always like to have more data, and I think any quality measure developer would love to test this out over and over again just for maintenance and that. So I certainly would love to have more data and get additional data, but I do feel like we have a decent sample for this.

MEMBER THOMAS: Great, thank you.

MEMBER MORT: So just since this has come up a few times, the data elements that you're using either in the risk adjustment model or in the observed or expected calculation, are those all derived from staff data elements put in

the medical record or are any of those from coding sources, administrative data sources?

MS. DEUTSCH: Sure. So most of the -- like age, we actually, I think we used the claims data for that. But a lot of it comes from the chair assessment data, but we did use comorbidities from the claims data. In part, we don't know that, necessarily, every inpatient rehab facility codes the same way. And consistent with other, like, readmission measures, we went back to acute care. So if there was somebody who was listed as having diabetes in the acute care claims record, we accepted that, even if it wasn't maybe in the IRF claims record. So we did use claims data from the IRF stay, as well as the acute care stay, in addition to the assessment data.

MEMBER MORT: And that would be the way it would continue to be analyzed. You go back to the incident in admission that led to the IRF stay and you pull ICD-9 comorbidities from that database? That's --

MS. DEUTSCH: Yes, so ICD-10. Right.

MEMBER MORT: ICD-10. Excuse me, excuse me. So my comment about reliability, Peter, just is that the gameability, in my mind, is highest when it's totally relying on administrative data alone. And these data elements come from two sources. The fact that you go back to the incident and hospitalization to look for additional comorbidities suggests it's pretty thorough, but, regardless, it's the same data used for either the risk adjustment or the O-to-E calculation. So if there's bias in it, it's biased in either methodology is how I understand it.

CO-CHAIR PARTRIDGE: Okay. Any further? Becky? Sorry.

MEMBER BRADLEY: Just to help me understand it because I know there is a great interest in using this tool. How long does it take to get to the data pool that you would like to see to use for expected scores? I mean, is that years, is that months, to have enough data

collected to be able to provide the information for the predictive piece that seems to be missing?

MS. DEUTSCH: So if you're asking what the time frame is for the measure, each measure, there's a time frame, so what patients are included. So we have in our application 12 months of data would be used. Is that, is that what you were asking?

MEMBER BRADLEY: So it would take 12 months to build a database to have the predictive number, the predictive validity and reliability that --

MS. DEUTSCH: So we're proposing creating the quality measure at the facility level with 12 months of data, and that's consistent with other measures. Some measures 24 months.

CO-CHAIR STILLE: Okay. And for the testing, did you use like a retrospective data set from a few years ago or.

MS. DEUTSCH: The data were collected

between 2008 and 2010.

CO-CHAIR PARTRIDGE: If there's no further discussion.

MS. ALLEN: Voting on reliability: one high, two moderate, three low, four insufficient. Voting starts now. All votes are in: zero percent high, 77 percent moderate, zero percent low, 23 percent insufficient.

Voting on validity: one high, two moderate, three low, four insufficient. Voting starts now. All votes are in: 8 percent high, 69 percent moderate, 8 percent low, 15 percent insufficient.

CO-CHAIR PARTRIDGE: Moving on to feasability. Any discussion on this element? Sam?

MEMBER THOMAS: I just want to ask do you have any data on how long it takes to administer or to grade this instrument?

MS. DEUTSCH: So clinicians are assessing patients on these activities pretty typically. Some of the items are things like car

transfers that are done at best-practice facilities. I'm not sure that I could say 100 percent of IRFs are looking at things like car transfers at discharge, but the best-practices facilities recommended all of these items.

So the assessment is happening anyway in most cases, and so, in terms of, like, the documentation piece, I'd have to look up the specifics, but, I mean, there's basically taking the information and then thinking through what the score is and documenting it. And so there's seven self-care items and 15 mobility items.

MEMBER THOMAS: So any member of the rehab team, a therapist, nurse, physician, any member could conceivably do this grade and there's not a restriction on that.

MS. DEUTSCH: Correct.

CO-CHAIR PARTRIDGE: If there are no further comments on feasibility, we want to vote.

MS. ALLEN: Voting on feasibility: one high, two moderate, three low, four insufficient. Voting starts now. All votes are in: 46 percent

high, 38 percent moderate, 15 percent low, zero percent insufficient.

Voting on usability and use: one high, two moderate, three low, four insufficient information. Voting starts now. All votes are in: 46 percent high, 38 percent moderate, zero percent low, 15 percent insufficient.

Voting on overall suitability for endorsement of Measure 2634, change in mobility score for medical rehabilitation patients, and Measure 2636, discharge mobility score for medical rehabilitation patients: one yes, two no. Voting starts now. All votes are in: 85 percent yes, 15 percent no.

CO-CHAIR PARTRIDGE: Take a deep breath. Move on to our last measure, and our friends to our left are up again.

MS. DEUTSCH: All right. Thank you. Okay, great. So this last measure is change in mobility for patients admitted to a long-term care hospital on a ventilator. So just in terms of the background, I mentioned, I think

yesterday, that long-term care hospital patients are very, very sick and they are often having limitations in mobility, as well as self care, mobility in particular. There are risks for additional limitations to develop because somebody is basically on a ventilator and not able to move around a lot sometimes.

So this measure is a risk-adjusted change in mobility measure between admission and discharge. And, again, it only applies to long-term care hospitals who are admitted on a ventilator.

We looked at our PAC PRD data and selected items that we thought made sense based on that data, what's feasible to collect in a long-term care hospital for these very sick patients. We presented this to our technical expert panel, so we only have eight mobility items just because it's not feasible to expect people to go up and down the stairs, for example. So we have bed mobility items. So we have some low-functioning, items that work for patients

with low function, as well as transfer items, so important things like getting on and off a toilet, as well as up to walking. This is consistent with items that are reported in the literature for patients in long-term care hospitals. The period of time for the measure is 24 months just because it takes a longer time to have enough patients who are admitted with this particular condition.

So in terms of importance, I kind of mentioned that these patients are very high risk for having further decline in their function, so that's important. And also Congress has actually weighed in on this. There is law, the Bipartisan Budget Act, that was passed in 2013 that requires Medicare to establish this quality measure by 2016. And the reason that that is such an issue is that there's a lot of payment reform happening in the long-term care hospital, so it's important to have a counter in terms of measure and quality when there's, again, pressures related to payment reform.

We have several exclusion criteria. This was based on input from the expert panel, as well as our public comment, so it's similar to measures we presented before. Incomplete stays, we only included people 21 years and older. We have a risk-adjustment model that we presented, and, again, we used generalized linear models with general estimation equations. We calculate an observed over expected times the national average, so you get a risk-adjusted change score.

I think you've heard enough from me on other things, so, Stacy, do you want to talk about standardization?

MS. MANDL: Sure. So, again, this is Stacy from CMS. It's come up a couple of times, and I thought I would just use this as an opportunity between the two measures to sort of touch on this concept of uniformity and standardization at the data element level.

So here's a great example where a law was passed that requires a measure. Data elements and the importance of having data elements that

are standardized, especially at the assessment level, really helps to make sure that you have apples to apples. I get it. It's very critical to have the trainings in place and all of the iterative trainings in place and so forth because of that, but it's far better to have that than to have a measure that doesn't specify the data element level to the level of detail that we do.

I just wanted to touch on that. And it's such an important concept that Congress also just passed another law called the IMPACT Act that requires that for multiple purposes, not just for quality reporting but also for the purposes of transferability of information, interoperability.

So I just wanted to just sort of touch on that why, yes, we're hearing you loud and clear at facility, but at the data element level there has been testing. So I'll hand it back over.

CO-CHAIR STILLE: And I'm the discussant for this one, as well. I think,

overall, I think it's a very interesting measure. I appreciate the discussion of law in need of a measure because my primary quibble is we don't have a whole lot of data yet. You know, it's 103 patients, basically, to kind of look for -- am I look at the --

MS. DEUTSCH: Four hundred and fifty-five.

CO-CHAIR STILLE: Okay. So the 103 must have been a subset. So a still relatively small number, but not that many patients need ventilator support, so it was probably hard to get a sample that was big enough. So we need to get more data as times goes on with that.

I had a couple of comments and sort of worries about exclusions. One, I need some input from the commissions that are used to dealing with adults. But it seems maybe inappropriate to exclude all progressive neurologic conditions, especially the ones that go up and down. MS and Parkinson's I'm wondering, in particular, you know, a lot of patients with MS and Parkinson's,

they do great, they do terrible, they do great, they do terrible. Even if they need to be on a ventilator, they may not need to be on a ventilator in six months to a year if they do better. And so I was sort of wondering about the reasons for some of those exclusions.

MS. DEUTSCH: Do you want me to address that or are you asking --

CO-CHAIR STILLE: If I could hear from another clinician, that would be good. And then -- yes. So maybe Liz?

MEMBER MORT: Well, the neurologic deterioration or conditions that lead to either fluctuations or deteriorations over the long term are highly variable. I'm hesitating. I can understand why they excluded that population. On the other hand, I wonder, of the 300,000 people on ventilators every year, what proportion of that --

CO-CHAIR STILLE: Exactly. That was my point.

MEMBER MORT: -- and it might be

worthwhile taking a look at the data to see how many patients, what proportion of the 300,000 ventilated patients annually fall under those categories and which categories. There might be some that you would leave excluded, like ALS for example, but others that you might include. So I think it's worth exploring with a bit more detail.

CO-CHAIR STILLE: Great.

CO-CHAIR PARTRIDGE: Becky?

MEMBER BRADLEY: In keeping with the idea of developing measures that cut across all settings, I was just curious because there are people on ventilators at home and people on ventilators in nursing homes and in inpatient rehab, why was this one limited, especially since you're trying to collect data on as many patients as possible, why did you limit it?

CO-CHAIR STILLE: Go ahead.

MS. DEUTSCH: Okay. So I have a question from Chris and then went from Becky, so I'll deal with Becky's first and then we can come

back to the other.

So inpatient rehab facilities have very few patients overall, and it's a select group of IRFs that admit patients with ventilator. So I don't think we would get a big enough sample. So, I mean, if I could just say, you know, anybody on a ventilator, I'd love to include in this measure, great. But then facility-level analysis would be very difficult at the IRF level. It would be very difficult at home care. I just, there wouldn't be many home care agencies that would have a whole lot of data.

So long-term care hospitals, part of this law that was passed that I mentioned before, they basically say, you know, patients who are chronically critically ill, include patients admitted on a ventilator. And there's criteria within the law in terms of what chronically critically ill is, but ventilator is their prime example. So that's really where patients are at this point, in terms of the post-acute care

world. So does that help, Becky?

MEMBER BRADLEY: Well, it helps understand because I do recognize in inpatient rehab, you know, you don't have that many. But if you're looking at where they're best served and being able to take them from one level of care to another and share information and, you know, hands off providers, it seems like it would be useful if that is the intent of developing a core measure that cuts across all.

MS. DEUTSCH: So that's actually a great comment. So I think this goes to Stacy's comment. Standardizing assessment data cross all the post-acute care settings would allow us to do research studies to examine exactly this point and be able to risk adjust. So without the data, you just cannot, you know, do this at this point.

So this would really open up the door to be able to do research and then, you know, potentially develop quality measures, as appropriate. And then you'd be able to look at function, you know, our people who are home and

have, you know, on ventilators, do they actually do better? So great, great point.

MS. MANDL: CMS thanks the comment.

MS. DEUTSCH: So can I answer, Chris, your question, unless, Sam, did you also want to weigh in?

MEMBER BIERNER: I mean, I understand or I assume that your reason for excluding some of these was because some of them have a variable course and, for the purposes of creating this instrument, you know, you don't want to introduce unnecessary variability into your study. I would hope that, you know, over time, if this gets implemented, then one could use it and see what kind of data one gets from that population. But I assumed it was excluded mainly for the purposes of developing the instrument.

MS. DEUTSCH: Yes, that was exactly the point. In fact, we talked with our expert panel about these diagnoses in particular, and they said, well, if the person is admitted to an IRF they should be included in that measure

because IRFs should be focused on functional improvement. But if somebody is admitted to an LTCH they may or may not have necessarily the same prognosis, and so it's really unfair.

And to Peter's, I know, Peter, you have brought up several times potential unintended consequences. If you have patients who you know may not do well on a particular measure, they might be excluded from access to that level of care, which may be very appropriate. So that was really the rationale, and we did present data from our sample. We had 12 patients who had ALS, which was 1.59 percent. The multiple sclerosis/Parkinson's, it was fewer than 11. CMS data use agreement, we can't report raw data when it's less than 11, but a very small percentage.

CO-CHAIR STILLE: Okay. So, I mean, that just points out that it's really important to potentially change the measure if you're looking at it in different settings, you know. Include it in one group, exclude it in another

group, because they're different people.

MS. DEUTSCH: This is all about goals of care.

CO-CHAIR STILLE: Yes, yes, great. Okay. My only other question was I didn't see a whole lot of data, and maybe there aren't, about a performance gap at this point so.

MS. DEUTSCH: Yes, I mean, there's not a lot of literature about long-term care hospital patients, in particular ventilator patients. And I think our literature review speaks to the idea of early mobilization being a pretty new concept in providing care, and so it's just a really early thing and we're just trying to get ahead of the curve and make sure patients are getting the best care possible and functioning as independently as possible.

CO-CHAIR STILLE: Other comments on importance? Becky and Liz, can you put your things down? Thanks. Unless you had another point? Okay.

MS. ALLEN: Voting on evidence for

Measure 2632: one yes, two no. Voting starts now. All votes are in: 100 percent yes, zero percent no.

Voting on performance gap: one high, two moderate, three low, four insufficient. Voting starts now. All votes are in: 8 percent high, 62 percent moderate, zero percent low, 31 percent insufficient.

Voting on high priority: one high, two moderate, three low, four insufficient. Voting starts now. All votes are in: 23 percent high, 77 percent moderate, zero percent low, zero percent insufficient.

CO-CHAIR STILLE: Okay. So in terms of reliability and validity, again, it's a little bit tough because there's not a whole lot of data out there. So I want to rely on the psychometricians in the group to talk about whether it's enough for what we have right now.

There is item-level reliability for the items that feed into the overall assessment. There's not much data on the reliability on the

whole measure. And I'm a little bit out of my depth when it comes to talking more sophisticatedly than that. So, Dawn, please, thank you.

MEMBER DOWDING: I don't think I'm any more expert, but I just want to clarify because it's a bit unclear from the description. This isn't a change score. This is another one of the ones where you're calculating an expected functional score at discharge and comparing that to what the observed score actually is.

MS. DEUTSCH: So this is a risk-adjusted change score. It's mobility. So we calculate an observed change for the eight mobility items. So we look at the discharge score minus the admission score. So we calculate that as the observed. We apply the risk adjustment data to the patient-level data to calculate an expected. We then divide or create a ratio from observed over expected, and then that ratio is multiplied by the national average, which is similar to what we did with the IRF

change in mobility.

MEMBER DOWDING: Yes. And I guess my question is, given the small population on which you have got data for and given the, unless I'm reasoning it wrong, your predictive model only predicts 26 percent of the variant, so there's a huge amount of noise in that model, why have you gone for that very complicated calculation for this score? Why haven't you gone for a simpler change in function score? Because I guess I'm really supportive of this, and I understand that you don't have a huge amount of data. But given all of that, I would have expected to see a less complicated calculation than a more complicated one. And, also, could you also give us some ideas about the number of patients that we use to develop the predictive score? Sorry. That's a lot of.

MS. DEUTSCH: So our sample was 455 for the model. So our observed over expected times the national average is the same that we had presented for IRF. I guess it's a pretty

standard thing. I mean, I'm an epidemiologist. It's an epidemiology 101 risk adjustment. It really is important to risk adjust. We tried all kinds of comorbidities. We actually tried every HCC or groups of HCCs that we could. And, again, we went back to acute care, so I think -- if you have that question.

So, I mean, we tried our best to risk adjust, and I guess I feel like it is critical to risk adjust. So I guess --

MEMBER DOWDING: I mean, I'm not an expert in this, but there are other ways of risk adjusting which don't involve a predictive score. I mean, it's this whole business about expected versus observed. You can risk adjust the change without having an expected score that you then calculate. I mean, I'm not an expert, but we've seen other measures where they've done risk adjustment which doesn't involve this expected --

MS. DEUTSCH: I think all of the measures we've heard about today used observed and expected.

CO-CHAIR STILLE: So it's risk adjusted, as opposed to proportion passing a certain expected, difference with expected, so I think the number is a little different then.

MEMBER BRADLEY: Well, I guess, if we're looking at a quality measure and a performance measure and we were kind of shooting in the dark with an expected, how will that expected score be used for performance improvement or in a facility level?

MS. DEUTSCH: You're asking how it would be used, the performance one? So the facilities could get a report and realize that they are different than a national benchmark, so they would get their facility data and there would be national benchmark data ranges, and so they would get a sense about whether they're performing better than the national average, worse than the national average, and so they could respond to that, obviously, if they think they have room for improvement.

MEMBER BRADLEY: Well, and I guess

that's where, not being a statistician, I would be concerned because if we're not certain that that predictive score is reliable, we may be spending a lot of resources trying to adjust on something that hasn't really been fully tested for reliability from the predictive standpoint.

MS. DEUTSCH: So I think, I mean, our score was about 26 percent. I guess I don't know if NQF staff can comment about, relative to other measures, how that -- because, I mean, I've definitely seen worse.

MEMBER BIERNER: Well, I mean, compared to just, maybe not just specifically other measures, but that would mean the lower end of the, you know, lower end, beginning of moderate level, for R-squared value. But I would assume that, because you have a relatively small sample size, that, over time, you would revise this regression model based on your data. So then we'd expect that we'd get updated as new data comes in, and every year or two years or something it would be updated.

But in terms of how it would be used, I would assume that it would allow facilities to compare their population and their performance to others in their region or nationally. And so it would become more useful over time as more data is accumulated.

MS. DEUTSCH: Yes. And this was finalized in the Long-Term Care Quality Reporting Program this August 2014, so this is being implemented April 1, 2016. So there will be more data available, and I guess, in terms of R-squared, if you had a really, really high R-squared, that means that you're explaining all the variants. And so, you know, we think there's facility input that makes a difference here. So we definitely would love to do a whole lot more and try to get a higher R-squared, but we did try everything that we got in terms of suggestions from public comment, as well as our expert panel.

MS. MANDL: I just want to add there's only 400 LTCHs in the whole United States, if that matters.

MEMBER MORT: Since we're talking about the O-to-E, 400 patients in your sample size, obviously, is small. And as you revise it, you'll have more patients in that sample. I had a comment about a slightly different aspect of it, which is mobilizing patients on vents is a relatively new area of focus clinically, so expected based on current practice is not where we necessarily want to be. So I think you want to beware of having people focus on a goal that probably isn't aspirational, and that doesn't necessarily relate to some of the other things that we've been working on for a long time. We have a better level of understanding about where we can expect patients to get. Do you know what I mean?

MS. DEUTSCH: Yes. So this is absolutely an area that's evolving, and that's part of -- I think NQF has maintenance endorsement renewal and all that that's needed, and I think CMS is looking at measures, every single measure every year to make sure that it's

keeping up with the evidence. So absolutely.

CO-CHAIR PARTRIDGE: Ready to vote?

MS. ALLEN: Voting on reliability: one high, two moderate, three low, four insufficient. Voting starts now. All votes are in: 8 percent high, 54 percent moderate, 23 percent low, 15 percent insufficient.

Voting on validity: one high, two moderate, three low, four insufficient. Voting starts now. All votes are in: zero percent high, 69 percent moderate, 15 percent low, 15 percent insufficient.

CO-CHAIR PARTRIDGE: Discussion on feasibility? Liz, is that a card up or --

MEMBER MORT: No, I'm sorry.

CO-CHAIR PARTRIDGE: All right. I think -- Nadine?

MS. ALLEN: Voting on feasibility: one high, two moderate, three low, four insufficient. Voting starts now. All votes are in: 23 percent high, 69 percent moderate, 8 percent low, zero percent insufficient.

Voting on usability and use: one high, two moderate, three low, four insufficient information. Voting starts now. All votes are in: 23 percent high, 54 percent moderate, 8 percent low, 15 percent insufficient information.

Voting on overall suitability for endorsement of Measure 2632, change in mobility among patients requiring ventilator support: one yes, two no. Voting starts now. All votes are in: 92 percent yes, 8 percent no.

CO-CHAIR STILLE: Great job, everyone.

(Applause.)

CO-CHAIR PARTRIDGE: And our deep thanks to the team at CMS and RDI. We appreciate your patience, your willingness to sit through two days. Peter has some comments.

MEMBER THOMAS: I just want to underscore what Lee said at the very beginning of this yesterday morning with the comparison between the staff preparation of the measures last time around and this time around and pay a huge compliment to the staff of the NQF for

putting together the materials in the way you did and the staff review and eliminating the prep call and instead focusing on the way you did it this time. We could have never have gotten through this 28 measures if it were done the other way, and so I strongly endorse sticking with this approach, even though it's a lot more work for you, I know, in going through it. But, boy, it just made this -- this is a long process to get through 28 measures as it was, but it made it so much easier. And I just wanted to say thank you very much.

(Applause.)

CO-CHAIR PARTRIDGE: We're going to open the mic for public comment. I'm going to turn first to the people in the room behind me. Please, just introduce yourself and then go ahead.

MS. HART CHAMBERS: So I think we have it now. Okay, thanks. So I'm Jayne Hart Chambers with the Federation of American Hospitals. I thank you all for your hard work

over the last two days. I was unable to be here in person yesterday.

But this is a challenge looking at this particular set of measures, and the discussion that Peter raised about the difference between public reporting and accountability programs and what's considered a value-based payment program on public reporting. It's going to be very difficult if some of these measures are deployed in different ways. And to have a measure deployed in the IRF Quality Reporting Program where there still is the opportunity if you don't report appropriately or get all your data in on time or whatever to be penalized and have a different measure in a pay-for-performance program will create a lot of confusion. And I just encourage, as we go forward, to try to find ways to create measures so that we have a smaller measure set that's more simplified that can get at the issues that are really important to both patients, providers, and improving care. I mean, ultimately, what we're trying to do with quality

measures is improve care.

So, you know, we're making steps. But I am very concerned that we still have a lot more to do. Thank you.

CO-CHAIR PARTRIDGE: Thank you, Jayne. On the phone?

OPERATOR: If you'd like to make a public comment, please press star and then the number one. There are no public comments on the phone line.

CO-CHAIR PARTRIDGE: Okay, thank you. As we all know, we did not have a chance to discuss today related and competing measures. We will do that, and staff will be back in touch with us about when.

For the balance of this day, we have a little bit more than an hour. We propose to go into executive session, have a chance to debrief from these two days of work, chat a little bit about our staff here at NQF about ways that we think this process is terrific and ways that we think it might be improved. And then we'll go

home by subway.

So thank you all for coming behind us, and we'll stretch a little bit and then regroup at 2:00 and say goodbye to Chris and hope he makes it to the airport.

DR. BURSTIN: And a special thanks to Chris and Lee for really leading what was a pretty difficult couple of days. So thank you. Great chairing.

(Applause.)

(Whereupon, the above-entitled matter went off the record at 1:54 p.m.)

N

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: 01-22-15

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