**NATIONAL QUALITY FORUM**

**Moderator: Kim Patterson**
**February 12, 2020**
**12:00 pm ET**

Sam Stolpe: Hello and welcome everyone. This is the Patient Experience and Function Measure Evaluation Web Meeting for the fall 2019 cycle. This is Sam Stolpe speaking and I'm delighted to welcome you.

Now we have two measures that we're going to be considering over this call. And why don't we just do a quick check in to make sure that we have our co-chairs on the line?

Chris Stille: Yes. Good morning or good afternoon. It's Chris Stille. I'm here.

Sam Stolpe: Okay, very good.

Lee Partridge: Good morning Sam, this is Lee Partridge.

Sam Stolpe: Good morning.

Gerri Lamb: And good morning, this is Gerri Lamb.

Sam Stolpe: All right. Our dynamic trio, so thank you, I'll have for the three of you for both being on the call for your participation and your leadership.

We're going to have two measures that we're evaluating as I mentioned, but because we only have two hours that does put some time constraints on how we approach this meeting. We haven't scheduled an additional meeting, so this is the preferred do or die for us. So we're going to encourage you to have a robust discussion of course, but we do have some limitations on how much time we can save.

We're - I'm going to move forward with this - a brief introduction of staffs. So we have myself, Sam Stolpe, Suzanne Theberge, Oroma Igwe and Tatiana Munoz.

So I wanted to hand it over to our circulars. I know we have you on this call, but did you want the operating orders to welcome to our committee members?

Gerri Lamb: Sure. This is Gerri. Sam, you asked for the co-chairs, right?

Sam Stolpe: Yes, I do.

Gerri Lamb: Okay, yes. Just welcome everybody and it's good to be back together again and thank you all for reviewing the two measures we're going to be looking at today. And like Sam said, I'm really looking forward to a robust and thoughtful discussion, so I'm glad you're here.

Sam Stolpe: Okay.

Chris Stille: And this is Chris. I'll echo what Gerri said. Do we have any new members this cycle or anybody we should particularly say hi to?

Woman 1: Well, this is (Ranjana Hinder) and I am a new member.

Chris Stille: Great, welcome.

Woman 1: Thank you.

Tracy Wong: Hi. This is Tracy Wong. I'm also a new member.

Chris Stille: Welcome Tracy.

Tracy Wong: Thank you.

Lee Partridge: And this is Lee Partridge who picked up a cold in Sunny, California last week and will be very silent today. Welcome to our new members and to my colleagues.

Sam Stolpe: Thank you so much. So we are also joined today by Apryl Clark. She was NQF's Chief of Staff, currently serving as the Acting Vice President of the Quality Measurement Department. Apryl is going to go through a roll call as well as a - just inviting you to just briefly say which (unintelligible) organization you represent and any disclosures of interest that you might have.

I hand it over to Apryl.

Apryl Clark: Thank you Sam. So first before I start, I would just like to say a big thank you to all of you for joining us today. You volunteer time with us and we appreciate all of your expertise. We couldn't do our work without you and so I appreciate all the time that you give for us. So I'm sure we'll have a very good meeting today with the couple of measures review.

As Sam mentioned, we're going to combine introductions with disclosures of interest. So you should have received a disclosure of interest form specific to measure specific, the measure that we're going to (unintelligible) today.

In the form we asked you a number of questions about your professional activities. Today we'll ask you to orally disclose any information you provided that you believe is relevant to the committee. We are especially interested in grant research or consulting related to this committee's work.

Just a couple of reminders, you sit on this group as an individual, you do not represent the interest of your employer or anyone who may have nominated you for this committee. We are interested in your disclosures for both paid and unpaid activities that are relevant to the (unintelligible) of you.

And finally, just because you disclosed it does not mean that you have a complex of interest. We do all disclosures in the spirit of openness and transparency.

So I'll ask you to state your name, who you're with and anything disclosed and I'll start with our co-chairs. Gerri Lamb?

Gerri Lamb: Yes, I'm here and no disclosures relating to the two measures.

Apryl Clark: Okay. Do you want to just say who you're with?

Gerri Lamb: I'm here actually - well, I'm with Arizona State University, but I'm here as expert and care coordination and co-chair.

Apryl Clark: Great. Lee Partridge?

Lee Partridge: Lee Partridge. I'm a Senior Fellow at the United Hospital Fund at New York City and I have no conflicts of interest with these measures.

Apryl Clark: And then Chris Stille?

Chris Stille: Hi. I'm Chris Stille. I'm with the University of Colorado. I'm a General Pediatrician. I have an academic interest in measures and measure development for children and youth with special healthcare needs that are not actively doing anything like that and certainly nothing related to the measures under discussion today.

Apryl Clark: Great. Now, we'll move to our committee members. Richard Antonelli?

Richard Antonelli: Yes, hello. Richard Antonelli, Boston Children Hospital, Department of Accountable Care and Clinical Integration. I have no disclosures relevant to the measures today.

Apryl Clark: Great. Adrienne Boissy?

Adrienne Boissy: Hi there. It's Adrienne Boissy. I'm at the Cleveland Clinic, Chief Experience Officer. I have no disclosures to the measures today.

Apryl Clark: Great. Don Casey?

Don Casey: Hey, it's Don Casey. Can you hear me?

Apryl Clark: Yes.

Don Casey: Great. Hey, I'm Don Casey, President of ACMQ, the American College of Medical Quality and I don't have any relationships with industry or disclosures. I will say actually as a patient I've actually had direct experience many, many times with FOTO. So I'll just leave it at that.

Apryl Clark: Ariel Cole?

Ariel Cole: Yes. I'm a Geriatrician employed by AdventHealth in Orlando and I have no disclosures related to these measures.

Apryl Clark: Ryan Coller?

Ryan Coller: Hi. Ryan Coller, Pediatrician, University of Wisconsin, no disclosures.

Apryl Clark: Sharon Cross?

Sharon Cross: Hi Sharon Cross. I'm a Director in Patient Experience for Ohio State University of Wexner Medical Center. No disclosures or conflicts of interest today.

Apryl Clark: Shari Erickson?

Shari Erickson: Hi, can you hear me?

Apryl Clark: Yes.

Shari Erickson: Right. This is Shari Erickson. I'm Vice President of Governmental Affairs and Medical Practice at the American College of Physicians and I have no disclosures.

Apryl Clark: Great. Chris Dezii?

Chris Dezii: There you go. I was worried. I thought you skipped me here.

Apryl Clark: No.

Chris Dezii: Hi, Chris Dezii - no problem. Lead, Healthcare Quality and Performance Measures at Bristol-Myers. We have no disclosures relative to the measures to be discussed. Thank you.

Apryl Clark: Great. Dawn Hohl? Sherrie Kaplan? Brenda Leath?

Brenda Leath: Good morning - good afternoon. This is Brenda Leath. I'm President and CEO of Leath & Associates and I am the Certification Director of the Pathways Community Hub Institute for PCHI. And I do not have any disclosures or conflicts of interest related to these measures.

Apryl Clark: Great. Brian Lindberg?

Brian Lindberg: Hi, good afternoon. I'm Brian Lindberg, Consumer Coalition for Quality Healthcare in Washington D.C. No disclosures or conflicts.

Apryl Clark: Great. Ann Monroe? Lisa Morisse?

Lisa Morisse: Hi. I'm Lisa Morisse. I'm the Executive Director of Consumers Advancing Patient Safety and I'm a Patient Advocate. I have no disclosures. Thank you.

Apryl Clark: Great. Randi Oster?

Randi Oster: Hi. I'm Randi Oster. I'm the President of Help Me Health and I have no disclosures or conflicts of interest.

Apryl Clark: Great. So we just remind folks that to put your earphone on mute if you are not talking. We have a little bit of background noise. Charissa Pacella?

Charissa Pacella: Hi. I'm Charissa Pacella. I'm with University of Pittsburgh Medical Center. I'm the Chief of Emergency Services there. And I have no conflict of interest to disclose.

Apryl Clark: Great. Lenard Parisi?

Lenard Parisi: Hi. I'm an Independent Quality Consultant and a Lecturer of Thomas Jefferson College of Population Health and I have no disclosure.

Apryl Clark: Great. Deb Saliba?

Deb Saliba: Hi. I'm Deb Saliba. I am a Director at the UCLA Bourn Center for Gerontologic Research. I'm an Internist with board certification in geriatric medicine. I also worked part time at the Veteran's Administration where I see patients and also at the RAND Corporation. And I have no disclosures specific to these measures.

Apryl Clark: Great. Lisa Suter?

Lisa Suter: Hi. I'm Lisa Suter. I'm an Internist and Rheumatologist at Yale. I'm an Associate Professor. I see patients both at Yale and at our affiliate VA hospital. I also am Director of Quality Measurement Programs at the Center for Outcomes Research and Evaluation where I do develop outcome measures

under contract to Medicare and voluntarily with the American College of Rheumatology.

I am involved in patient reported outcome measures in rheumatology and orthopedics in those roles none of which deal with back pain nor do they conflict with FOTO. I do not have any direct back pain or FOTO related conflicts.

Apryl Clark: Okay, great. Peter Thomas?

Peter Thomas: Sorry, this is Peter. I'm with a law firm, Powers Pyles Sutter & Verville. It's a healthcare law firm. I'm a - I do rehab and disability law in policy and advocacy and I have no conflicts to disclose.

Apryl Clark: Tracy Wong?

Tracy Wong: Hi. I'm a Director of Quality Safety Value and Patient Experience at Seattle Cancer Care Alliance. And I have no disclosures.

Apryl Clark: Great. Is there anybody whose name I didn't call on the committee or whom may have joined after we started the roll call?

Great. So like to remind you that if you believe that you have a conflict of interest at any time during the meeting, please speak up. You may do so in real time during the call or you can send a message via chat to your chairs or to anyone on the NQF staff.

If you believe a fellow committee member may have a conflict of interest or behaving in a biased manner, you may point this out during the meeting, send a message to the chairs or to the NQF staff.

Do you have any questions or anything you like to discuss based on the disclosures made today? Great.

Don Casey: This is Don Casey. I am wondering given that NQF is a member organization and my assumption is that we're all representing our organizations. Is it - do we have to analyze whether or not the organization we're representing has a conflict? I actually have never thought of it this way, but it makes me wonder since, you know, I think we're representing member organizations as I recall?

Apryl Clark: So let me just clarify. You do represent as a multi-stakeholder our membership, but you sit on the committee as an individual. You do not sit on the committee as a representation of your organization or of anybody that may have nominated you. So we do individual disclosures rather than organizational disclosures. Does that make sense?

Don Casey: Yes. I would take it one step further. For example, I'm President of ACMQ. And so as an individual, I mean I don't have - we don't have any conflicts. I just - I'm trying to raise that bar a little bit higher. It's a question I think you should mull over. But I'll leave it at that for now. It just - it makes me wonder about the people - other affiliations, that's all.

Apryl Clark: Okay. You're always looking for sort of opportunities to things that are conflicts of interest and make improvements. So, you know, we'll definitely take that back. It is a great feedback.

Sam Stolpe: Thanks Don. Okay. Let me turn it over to Suzanne to walk us through a couple of slides before we begin our discussion around our two measures today. Suzanne?

Suzanne Theberge: Great. Thanks Sam. So I'm just going to do a brief review of some details for everybody before we get started on the evaluation, so it's all fresh in your minds.

As you know, we are looking at two measures today. 0425: Functional Status Change for Patients with Low Back Impairments and 0291: Emergency Transfer Communication.

Prior to your evaluation, we have already got input from a number of different bodies in the process the Scientific Methods Panel looked at measure 0425, that's our team of scientific - statistical and methodological experts who evaluate the scientific acceptability criteria and make a recommendation to you the committee and 0425 as a complex measure goes to that one.

And also we have a common period that opens in December and is still open. And so, you know, we take in comments at this point prior to your review as well.

So as I said, we did have one measure go to the Methods Panel and it passed that evaluation. So it comes on to you with that recommendation included. It is an outcome measure, a patient reported outcome measure and that's why it's deemed complex and it was rated high on both reliability and validity by the SMP. 0291 is also a maintenance measure. However that one is a process measure, so it was revealed by staff, it didn't go to the Methods Panel.

I will just pause briefly for any questions on the Methods Panel before I talk about our process.

All right, hearing none, I just wanted to quickly remind everybody of the major endorsement criteria. It has been several months since you evaluated any measures. So we like to make sure that's at the top of your minds. We

are looking at first importance which measures whether things can - the goal is to measure those aspects of the greatest potential of driving improvements.

So what we're looking at, is there evidence for the measure, is there empirical evidence for it and is there a gap or is there still room for improvement. Both of those are must pass.

Then we look at reliability and validity. First one, reliability and then validity to make sure the measure is scientifically found. So those are all both must pass.

Feasibility, it looks at the burden of the measure, how easy it is to implement and collect the data and calculate the measure. And then usability and use, it must pass for these two measures as maintenance measures and that looks at whether measures are in use and then how the measure is being used to give feedback to users and then also how the measure is changed by any feedback received. And then finally, we would look at any related or competing measures (if need be).

So just quickly I know you are all familiar with the ground rules, just we hope you've reviewed the measures and we would ask that you keep your comments concise, because we do have a lot to get through today and that you do your best to indicate agreement without repeating what other folks have said and that you of course focus your recommendations on a measure evaluation criteria and if you have any questions about those, our staff are here to answer them.

So the process today, prior to each evaluation we'll have the measure developer give you a brief introduction to their measure, provide some details and then we have asked a couple of committee members to service as lead

discussions and they will briefly present the measures to the committee and kick off the discussion. They will be followed by the discussions and other (C) committee members that we've asked to be the kind of first respondent. And then the rest of the committee will discuss and you will vote on your recommendations.

We do have the developers on the call today to join us and answer any questions that you may have and we would ask the full committee to discuss and then vote on each of the criteria before moving on the next criteria.

So we do ask that you start with evidence and you just talk about the evidence and then you vote and then you start discuss gap, you just talk about performance gap and then vote on that and then move on to reliability, et cetera, et cetera.

I already went a bit over. So we've discussed this, but generally we found having some assigned to kick off the conversation and makes things go a little bit more smoothly, so we start with that.

So the voting process today, I did just go over the criteria, but we - I have one thing to flag here is that for reliability and validity of 0425 which did go to the Methods Panel, you have the option to either just accept the SMP votes and say yes, we agree with that or you have the option to discuss and then vote on your own recommendations.

Do know that the Methods Panel really just looks at the testing itself. They don't consider questions about risk adjustment. They don't consider whether the correct population is included or excluded. For the clinical details of the measure, we're really relying on you.

So if you have any concerns with that, that's the time to bring it up and we shouldn't just accept the SMP recommendation you should, you know, choose to discuss and adjudicate that yourself.

And then one other note is that if a measure fails on one of the must pass criteria, there is no further discussion or voting. We just stop the evaluation and the measure does not go forward as recommended.

So as you all know, we do need a quorum which we have achieved. By our account, I think we have 21 folks on the phone and we need 16 of our 24 members. So we're in good shape there. If at any point during the call you do need to leave early, please let us know, because if we lose quorum, we will need to stop voting and complete that via survey. So definitely let us know if you need to leave early. I know a couple of folks have already done that.

So for pass, a measure to be recommended or to pass the criteria, it must have greater than 60% of either high plus moderate or pass if that's the option on the voting criteria. And then for - it does not have - we do not recommend it less than 40% of the high and moderate votes.

We also have this zone called consensus is not reached or the grey zone as we call it internally and that's anything on the must pass criteria that - you know, between 40% and 60, inclusive of both 40% and 60% is considered that a consensus was not reached on that.

The committee would perceive and then would not make a recommendation on the final - vote on the final recommendation for endorsement, you will re-discuss and do that again at the post comment call. So we will kind of keep you posted on that if we get into that situation as we go.

So we did send out an email link. You will need to be on the webinar to see these slides, but you also need to be in the pole everywhere platform to vote. It's a separate link. If you have any trouble accessing that, let us know. We send it out right before the call. And voting is for committee members only. So if you're a developer or someone else on the phone, you won't have access. So that link is just for the committee.

So I'm going to pause here and turn it over to Tatiana to do a test of the voting software to make sure that everybody has access to it and everybody on the committee and that it is working for you all before we dive into the measure evaluation. So Tatiana, can you activate the voting?

Tatiana Munoz: Yes, of course. Hello everybody. I'm going to activate the test and it will be a yes and no question. It should now be active. You should be able to see on your screen and the question can be do you like broccoli, yes or no. This is more to discuss to make sure that the voting platform is working for all of you.

Randi Oster: This is Randi. The test question I got was just said test, yes or no, it did not ask about broccoli.

Tatiana Munoz: That's okay.

Sam Stolpe: That's kind of a bummer actually, but we did want to get a good poll as to whether or not you guys think it's listed. But if the test is passed, that's fine to us.

Richard Antonelli: And a copy of the poll responses, is that the collective results from the process or is it just seeing what I did?

Sam Stolpe: So just see what your actual vote was for now.

Richard Antonelli: Okay.

Sam Stolpe: 21. You can either raise your hand on the platform or just call out if you are still - if you need any assistance from staff.

Suzanne Theberge: Okay. So any questions before we start the evaluation?

Woman 1: Yes, this is (unintelligible). Were we supposed to see the results of everything or just that am I seeing that this stayed on the test question?

Sam Stolpe: No, we're…

Suzanne Theberge: Yes, we - go ahead.

Sam Stolpe: Go ahead Suzanne. Okay.

Suzanne Theberge: No, we don't.

Sam Stolpe: Okay. I got this one. So yes, we're not going to post the results for this one, we just want to make sure that you're able to submit.

Woman 1: Okay, thank you.

Sam Stolpe: That's good.

Chris Stille: Sam, how many total we have? 22?

Sam Stolpe: We have 21 in total.

Chris Stille: Okay.

Lenard Parisi: I have a question Suzanne.

Sam Stolpe: Yes.

Suzanne Theberge: Yes.

Lenard Parisi: If I send you an email, can you - will somebody, you know, accessing the email right now?

Suzanne Theberge: We will check - yes, we can check the emails.

Lenard Parisi: Okay.

Sam Stolpe: Yes. Just send it to the patient experience and function inbox and we'll be able to respond from there.

Lenard Parisi: Done.

Sam Stolpe: Thank you.

Suzanne Theberge: And if folks have questions, you can also submit them via the chat box and the staff will respond as quickly as we can.

Lenard Parisi: All right.

Gerri Lamb: Suzanne, this is Gerri. I have a question about the must pass criteria. Do we have an option on any of them to exempt?

Suzanne Theberge: That's a good question. We do have one and that is for evidence. If you feel that the evidence provided in - for a measure does not meet the criteria, but you believe that is because the evidence does not exist, not because the evidence simply was not put in the document or the evidence just in is very good.

But if the evidence - because nobody - maybe nobody has done research on this or maybe it's impossible to do a trial, a randomized controlled trial for some reason, if that is the case then you would vote insufficient. And then if enough people vote insufficient, the committee has the option to then vote on whether an exception to the criteria will be granted and the measure could go forward under an insufficient evidence with exception.

But again that's kind of a special case for measures that are not supported by evidence for a really solid reason, you know, not for cases where, you know, the evidence just isn't very good.

Gerri Lamb: Thank you.

Suzanne Theberge: Any other questions? All right, well, with that I will turn it over to Chris as our co-chair to facilitate this discussion and kick off the conversation on Measure 0425.

Chris Stille: Okay. It sounds good. So this is Measure 0425: Functional Status Change for Patients with Low Back Impairments. This measure developer is FOTO focused on therapeutic outcomes. We - do we have the developers on the line with us? (Diana) or (Daniel) or both?

(Diana): Hi. This is (Diana Haze) and (Daniel Roche) is on the line as well.

Chris Stille: Okay.

(Daniel): Hello, this is (Daniel). Can you hear me?

Chris Stille: Yes.

(Daniel): It sounds good.

Chris Stille: And okay. It has been a while. Do I ask the developers to say anything first or do I have to lead the discussion and start first or that's based on that?

Sam Stolpe: No, you got it right Chris. So we're going to lead off with a three to five-minute introduction by FOTO.

Chris Stille: Okay, great. So why don't you go ahead and talk for about three minutes or so?

(Diana): Thanks, this is (Diana). As you also described, this is a patient reported outcome, performance measure with a patient reported outcome measure, IRT based and (unintelligible) risk adjusted model as the basis of the measure.

I'd like to mention a couple of things to color our discussion today. First, I'd like to emphasize that there is a free public access, very nice and easy to use version only a click away. So it's on the Web site, easily accessible and the provider can put their patient on it to electronically complete the computer adapted test administered version followed by the risk adjustment question. So it automatically calculates not only the patient reported outcome measure, but additionally the risk adjusted components and it's also risk adjusted results easily and for free there.

There is also the short form that can be printed. For instance, if they're in a rural area or in a setting where they don't have the internet access, there is that manual paper, pen for option as well. Those are all provided for free.

Secondly, part of what - you know, it's important first and foremost that we have statistical reliability and validity. So all of the items, the functional questions and so on go through expensive testing that that you're aware of, also factors that color our decisions on ongoing development of the measure often include the experience of the patients and the provider.

So some of the comments that we saw where related to - we think more about the patient and the provider's experience such as functional activities being more general and such, so we take both in to account, but I think it's important to recognize that sometimes decisions are made after the statistical issues being settled first and we also take into account experience.

Chris Stille: Okay, great, thank you. Again, this is a maintenance measure. It was reviewed by the Scientific Methods Panel at the high reliability and high validity. I will turn it over to the lead discussion who is Deb Saliba.

Deb, if you could just briefly summarize the community evaluation comments and then take it away with your own thoughts and then we will have Lisa Morisse and Lisa Suter chime in with their comments.

Deb Saliba: Great, thank you. So this NQF 0425: Functional Status Change for Patients with Low Back Impairment. As you just heard from (Diana), it's a patient reported outcome performance measure that is risk adjusted and if reports of change in functional status for patients age 14 years and older with low back impairment.

It can be used as a performance measure at the patient individual clinician and clinical level. And the developers present ample data at all three of those levels that we'll talk about when we get to some of the reliability and validity discussions.

The scores are reported on a scale of 0 to 100, a continuous scale, the higher the score is the better the functional status. And it mapped to the international classification that's functioning.

You know, it's not clear from this title and I don't know how well everyone, you know, sort of knows about this system. But the focus here is on patients that are receiving therapy. And obviously our focus on therapeutic outcomes a developer helps us understand that a little bit better.

It's under - as Chris said, its undergoing maintenance evaluation. It was originally endorsed in 2008 and then most recently endorsed again in 2015. Regarding the - I was asked to start after doing a summary of the measure with evidence.

This is clearly an important clinical problem in my practice and most practices it's also something we all know about in daily lives. And the ability to function as (Diana) talked about is really an important patient outcome.

The developers did present some new evidence since the last time the measure was endorsed for maintenance that links the outcome to a healthcare process. Typically they did an analysis on 2015 data that showed a relationship between the patient receiving an early interim assessment that not required as part of the measure, but a process that might indicate a more intimate attentive clinical environment. And that - there is relationship between that and

improvements of 2.5 points on the functional status outcome at the provider level.

And regarding evidence, 10 pre-meeting (comment core) submitted related to the measure evidence is important and eight were supportive of the measure, one saw there was insufficient evidence presented and one question the clinical significance is that 2.5 points change that I just mentioned above.

As this is a maintenance outcome measure for which the evidence of the link between outcome and rehab process have already been determined in prior reviews, the determination of sufficient evidence for this measure would not rely totally on this new evidence, so rather it provides additional support.

The preliminary ratings from National Quality Forum's staff is a path for evidence and I agree with that. I don't know if anyone wanted to have - had questions or wanted to discuss before we - do we vote? I can't remember, do we vote at this point or do we move on to opportunity for improvement?

Chris Stille: Right. I think Deb that you provide any comments that you had on the other criteria and then Lisa and Lisa also do - is that right NQF staff?

Sam Stolpe: So for now I'm just stay focused on evidence before we move to…

((Crosstalk))

Chris Stille: So evidence, evidence, evidence and then reliability, okay, great.

Sam Stolpe: Correct. Chris if it will be helpful, I can offer a brief explanation on the difference between the evidence requirement that NQF has for outcomes versus process in immediate outcomes and standard measures.

Chris Stille: Excellent, great.

Sam Stolpe: Okay. I know we've talked about this in the past, but I want to just make sure this is top of mind for everybody and considering this measure and what the NQF criteria is.

For outcomes measures, we do not require what is in terms of QQC, the Quality Quantity and Consistency analysis of a - essentially a systematic review of the existing literature and evidence that supports or may not support a measure of their consideration and the measure focus area.

For outcomes measure, we had measures that only require that the developer demonstrates that the structure, process, intervention or service that could be introduced to help improve performance on an outcome. We're not questioning whether or not the outcome is desirable, nobody has to prove that, we're assuming that.

And the only thing that we would want to show is that there is something that the accountable entity can do to try to move the needle and - or effectively move the needle so to speak by introducing some sort of process that would lead to improvement in staff estimation is that the interim testing using the same survey instrument.

Normally the calculation of this measure requires two tests, one at the beginning and one at the end of therapy and the residual is it serves as the basis for the measure.

The developer tested having an interim assessment as a - that's something that you could do to improve performance and in our estimation, that in fact check that box (unintelligible).

Chris Stille: Great, thanks. Okay. So Lisa or Lisa, any comments that you would add on evidence?

Lisa Morrise: Well, this is Lisa and I think Deb did a great job. I thought it was important for the initial assessment to occur and then the follow-on assessment. I feel like it gives - there is evidence that it gives some direction to the plan of care and makes the treatment therefore better in the long run resulting in better outcomes.

Chris Stille: Great.

Lisa Suter: Yes. This is Lisa Suter. I agree with both of - I don't have anything to add. I think the evidence is in support of this measure.

Chris Stille: Great, okay. I'll open it up to the rest of the committee then. Any comments or questions on evidence for this?

Dawn Hohl: I just wanted to let folks know I dialed in late, this is Dawn Hohl on the line.

Chris Stille: Hi Dawn, welcome.

Dawn Hohl: Thanks.

Sam Stolpe: Yes. Hi Dawn, this is Sam Stolpe with NQF. Could we ask you just to tell us what organization you're with and also any disclosures of interest?

Dawn Hohl: Sure. With Johns Hopkins Home Care Group and I have no conflicts to disclose.

Sam Stolpe: Wonderful. Welcome, Dawn.

Dawn Hohl: Thank you.

Chris Stille: Okay. Any other questions or comments for the committee on evidence before we vote on evidence?

Don Casey: Chris, Don. Are we looking at page 3 and 4? Is that the summary of evidence that is - I mean I honestly didn't scan the 135 page documentation, but is this the - it looks like this is the evidence on page 3 and 4. I mean there is some prelude in the first couple of pages, but that's what we're looking for, right? That's what we're looking at as far as what the evidence?

Chris Stille: Yes, that's what I looked for the most.

Don Casey: Okay. So to Sam's point, I mean - and I will tell you, I have had extensive experience as a patient with this. I probably - I can't tell you the number of times, I would say at least 50. If - what isn't coming through and what was provided here was to Sam's point the structure process service, et cetera that is leading to the change of 2.5. It looks like all that we have here is that they did, you know, baseline and then follow up within a period of time is the only intervention described.

So I'm - from the standpoint of the way I would look at this, it's not clear to me here, you know, when they say to demonstrate evidence of a structure process intervention or service that can influence the (outflow of) interest. The only thing we know is that they went to a physical therapist twice and that

they did the survey twice, that's it. We don't have - at least I couldn't find any other information on this.

So from the standpoint of looking at the 2.5 change in points, you know, we're going to - I guess we're going to get into the populations. I don't know if it's here or there, but 2.5 points at discharge which also isn't clearly defined. I'm confused about this and I would personally think that the quality of evidence is actually very low from a quality standpoint.

Now, I'm - you know, Sam knows I'm a hard driver on this, but I'm struggling with this a lot.

Sam Stolpe: Right and…

Deb Saliba: Yes. And can I add? This is Deb. So let me - and maybe I mentioned it, but maybe I didn't emphasize this enough. It's the maintenance measure. So in the last two times that it was - in the initial time that it was reviewed by the panel, when it's initially accepted, they went through the evidence that's linked - that shows that there is a structure process that are intervention which is the process that actually improved outcome.

So the evidence at baseline back in 2008 of the relationship between physical therapy and movement and improvement in back impairments was already debated and discussed and found to be sufficient by the panel.

What we're asked to look at is to just - if the developer brings us any additional information to say, you know, to note that, but not so much to readjudicate the original link. It's my understanding from the copious instructions that I got for this - for doing this discussion.

So as we - you know, we're not here to readjudicate that baseline evidence about the link between therapy or between knowing someone's function as a therapist and having better outcomes.

Don Casey: Okay. Well, I understand that and it makes sense. But I guess for the purpose of this part where we're asked to vote, can you summarize what the structure process intervention and service was briefly in terms of what came out of the initial evaluation? I'm just trying to get my arms around it and a little bit surprised it isn't included here even though I know, you know, you mold over and I think it'll be useful to know that.

Deb Saliba: Yes. And, you know, they don't give us all of the discussion from 2008. So I can hand it over to the developer if they want to…

Don Casey: Yes, just summarize.

Deb Saliba: …give a little bit, yes.

Don Casey: Real quickly, you know.

Deb Saliba: Yes. Let me let the developer take a stab at that.

Chris Stille: Great, okay. Let's do that. Thank you.

(Daniel): Hello. This is (Daniel). I like maybe to offer a clarification comment on this additional evidence and the 2.5 additional points you mentioned, because I think there might be a misconception (unintelligible) following the comments I just heard.

So first of all, the detailed description of this process, there is a document here, the document that you sent us for the preliminary (unintelligible). Not only the summary that we looked at the beginning of the documents and I think that's on page - that was taken 25, it goes on to two or three pages.

And the 2.5 points mentioned are not the difference between the intake and the discharge. We actually assessed an interim patient reported outcome, so one that's mentioned before is not necessarily in order to (continue) the measure, but one that's taken during the episode of care and we looked at that as a process. And the reason we did that is because it's not - this is - it's not a necessity for a therapist to do that.

They usually do that because they want to use the data for clinical decision making. We want to know how the patients are doing during the episode of care. And what we discovered is that clinicians that attempt to do that more than others and especially earlier on the episode of care get an additional 2.5 functional status point change at discharge. So it's a difference in change scores, it's not a difference between intake and discharge.

And this evidence was actually published in a manuscript not long ago, so - and we've referenced this paper so we can obviously look at more details. But I just wanted to clarify that point.

Chris Stille: Okay. Great and then Deb, did you have any other questions about the original evidence that you wanted the developers to answer?

Deb Saliba: No.

Chris Stille: Okay, great.

Deb Saliba: I do not. I was just, you know, giving them an opportunity to respond to the questions from the panel.

Chris Stille: Okay, that's great. Yes, the only other thing that - this is Chris. The only other thing that I noted is that in this summary during the last review in 2014, the measure was granted in exception to evidence because of lack of clear retrospective studies and the unethical nature of any (unintelligible) that you would do related to this. So I think that's just an important background for the committee to know.

Richard Antonelli: Chris, this is Rich Antonelli. Do you want us to put our hand up electronically or just jump in this way?

Chris Stille: Just jump in this way. I only have two screens and they're full of other things. So I can't see hand. And so staff, if there are hands that I need to know about, please let me know.

Richard Antonelli: All right. So I like to get myself in the queue. I don't know if anybody is ahead of me. I will be happy to wait my turn.

Chris Stille: Go ahead Rich.

Richard Antonelli: Okay, thank you. So I'm - I have the same confusion that Don did to the extent that we're getting presented evidence that there is a movement and especially the linkage to earlier assessment. That makes sense to me, but I don't know that in the context of considering this measure for maintenance reconsideration, I actually think it's confusing to present that, to see a bump when you do an earlier assessment that's great. That tells me that the measure is sensitive and may have some validity.

But I do think it's important for this committee to understand what we're being - at least my understanding what we're being asked for is this measure adds its stance, you know, is there appropriate evidence for that? So I - then the only other question that I had in - I've got other questions for other domains, but the other thing I tee up for the group here is I may have missed it, I actually did review all of the pages, but I may have missed it.

The evidence for lumping adolescences into this and I rarely at the NQF identify myself as a pediatrician but I will, because I've got a fair amount of clinical experience here.

What's the evidence that the sources of low back pain for adolescences are in fact amenable to the same measurements that the - probably a different set of pathologies for adults?

Deb Saliba: So this is Deb, can I jump in? I had the exact same question and because 14 did seem a bit odd and when I got to page 91 in the wee hours of the night, it - I did (unintelligible) to myself and it answered my questions regarding age less than 18. So if you want to look at page 91 about the age inclusion perhaps that could be helpful.

Richard Antonelli: So I did - so thank you for that and I actually did find this, but I guess what I'm trying to say at least as a clinician, this doesn't satiate my appetite to learn more.

Deb Saliba: Okay, that's fair. But I will say that they did look at the question in the 14 to 17-year old group and for other people that didn't read all the way to page 91. It did - they did look at the residuals for the patients 14 to 17 and then compared it to the 18 and older group and looked at correlations in both (unintelligible) and interclass correlation coefficient. And yes, so I felt

satisfied that they had at least addressed. I understand what you're saying too that it might be nice just to have those at some point be able to pull that data out.

And my understanding is as a clinician you can sue that with the program, with the software program you can look at subpopulations, but let me ask the developer if that's the case which is different from the measure.

Chris Stille: Great. If one of the developers want to chime in that will be helpful, thanks.

(Diana): (Daniel), do you want to speak to the scientific aspect and perhaps we could comment clinician-to-clinician about patients with low back pain being seen in physical therapy.

(Daniel): Yes. So just a quick comment, actually I already mentioned and that's from scientific perspective we really wanted to make sure that for those younger patients our risk adjustment model is doing a good job and after getting different results and we would respect that with other patients.

So that's the analysis that was just mentioned and we were satisfied with that. So we feel comfortable scoring the measure for those patients as well using the same risk adjustment model. Now age is one of the factors adjusted in the model, so that helped as well. In terms of comparisons between clinicians, maybe (Diana) do you want to respond to that?

(Diana): Yes. I guess one of the - I heard the gentleman say you use the - you're a pediatrician and differences in that population. So I just wanted to speak clinician-to-clinician as well. (Daniel) and I saw physical therapy patients for years and years and years including lots of adolescences with low back pain.

So with the physical therapy presentation it's similar, it's different, but the question is, is it different enough to justify a completely different measure and so on with that whole when we already have too many measures in healthcare.

And at the same time our last NQF experience is where such that the committees at the time were more - were suddenly encouraging more validation on younger age groups. So that was part of the reason that we really doubled down to make sure that this is appropriate for those age groups.

Richard Antonelli: So if I may - this is Rich Antonelli, the same pediatrician from Boston, I just sort of fall along that. So what I'm struggling with is not the use of this as an assessment tool. I - actually I'm thrilled that we have that and that's in the spirit of that. But what I'm thinking about and this is the term I learned there at the NQF over a decade ago is to measure a fit for purpose.

So what I'm concerned about is this measure could eventually find its way into an accountable care contract that looks at improvement scores on this measure and not really understanding in enough depth the risk adjustment methodology.

I totally get it as a clinician that getting earlier assessment likely will improve the outcomes, but I'm - I haven't seen enough information in the presentation today that says okay, I get it. We can do this for 14 and above, but if this migrates into an ACO contract -- for example -- that includes adults and kids or adults and pediatric patients, it could be problematic because I don't have enough insight into how you have risk stratified looking at the level I'd say performance across the clinician group. So that might be a later domain discussion. But I'm teeing this up with that so that notion of fitness for purpose.

Don Casey: Yes. Rich, this is Don. I have the same concern here, you know, with respect to the high degree of variation with the population denominators of the measures so to speak. You know, you got acute and chronic, I mean I've had - I'll just tell you. I've had low back pains since I was 16. So I will be excluded - would have been excluded then.

But I also have it now and social determinants, you know, et cetera, et cetera. Cancer isn't in here for example. There is just a lot of question. I'm not - it's not disputing the utility of using this, you know, in the heat at the moment with the patient. It's this issue of purpose that I'm having trouble with.

Chris Stille: Okay. So Rich and Don, this is Chris. I'm trying to understand your concerns. It sounds to me as though they're a little bit more in the way of validity among different subpopulations as oppose to evidence, right? I wonder if we should maybe…

Don Casey: Well, I think the two are related Chris.

Richard Antonelli: Yes.

Chris Stille: So far.

((Crosstalk))

Richard Antonelli: And Chris, that's what I'm saying. I'm very comfortable deferring the bulk of this conversation to later on, but because I viewed the presentation of evidence with all due respect, a bit in this place because I don't think it was evidence about the measure, I think it was evidence that early assessment makes the difference.

So right, and I'm not voting it down, I'm just pushing back on I'm not convinced that there is no evidence. Hard stop. Happy to move it later on as we talk about validity. I'm concerned about risk adjustment across ages and I'm really concerned about fitness for purpose.

Chris Stille: Okay. So yes, I'll be inclined to hold that thought, but yes, it does sound like it's more - the evidence is more to try and boost the validity unless others disagree.

Don Casey: I agree.

Chris Stille: All right. We do kind of have to start to move if we can. Any other major concerns about evidence before we'd go? Okay, all right. Let's vote on evidence then if it's time staff.

Woman 2: Voting is now open for evidence on Measure 0425.

Amy Moyer: Option A is pass, and Option B is do not pass.

Man 1: Okay, we're at 19 votes now. We're looking to get to 21.

(Peter Thomas): This is (Peter Thomas). All you need to do is press the button and it turns blue and you're done, right? You don't need to (unintelligible)…

Man 1: That's it. Thank you, (Peter). Has anyone who wants to vote not voted? We're at 20 now.

(Jeri Lam): Yes, this is (Jeri Lam). It's not loading for me.

Man 1: Okay, (Jeri), if you would like to send it via chat or give your vote verbally…

(Jeri Lam): Just did.

Man 1: All right, thank you very much.

Amy Moyer: Voting is now closed. We have a total of 21 votes with 19 being pass and two being do not pass. Therefore Measure 0425 passes on evidence.

Man 1: Very good (unintelligible).

(Deb): Now I'll talk about opportunity for improvement.

Man 1: Yes, please.

(Deb): So again, this is a maintenance measure. And the developer provides data that shows variation in performance that both the clinician and the clinic level when it's grouped into three categories. And the low group of 18% of the clinicians are in the low, 58% in average, and 14% in high, and were distributed at the clinical level with the clinic level low being 29%, average 52%, and high 19%.

Similarly, when they present the data - again, they presented a lot of supporting data. And when they present the data (unintelligible) in different sample years, you also saw a distribution - again supporting the idea that there are opportunities for improvement. And the preliminary staff rating for the opportunity for improvement was high.

Of the ten pre-meeting comments one reviewer questioned how categories of high, average and low considered the standard error of measurement. And another reviewer felt that sufficient or moderate evidence of performance gap

were opportunities - everyone else -- I'm sorry, everyone else -- thought that there were sufficient or moderate evidence of performance gap and therefore opportunity for improvement.

I don't know if we want the developers to speak to this one review that asks about the standard error of measurement.

Man 2: Yes maybe just briefly to clarify.

(Deb): Yes.

(Daniel): Yes, this is (Daniel). I'm happy to do that. So the standard error of measurement is used to calculate the measure's reliability. Once we've done that and we think the measure is reliable enough -- meaning the error term is not too large to make the measure inaccurate -- then we go on to other testing. And the performance gap is one of those additional testing.

So when you look at the difference in three quality levels using the average residuals per provider in the 95 confidence intervals associated with these averages, that's how we look at those differences. So we want three distinct groups. So I think the standard error of measurement comes into play much earlier in the game, when we the measure. Not so much at this stage when we have determined already that yes, (unintelligible).

Man 2: Great, thank you. That was great. Okay.

(Deb): So can I ask - am I supposed to go on to talk about disparities in this section? Or do we want to - and then have (Lisa) comment? Or do want to just let (Lisa) comment now about opportunities for improvement?

Man 2: I think disparities are more related to validity, if I get that right.

Man 3: No, they're actually included in this section here.

Man 2: Oh, are they? Okay.

Man 3: Yes, please discuss (unintelligible).

Man 2: Okay, go ahead and talk about that, then.

(Deb): Okay, so I will throw - talk about disparities and then ask my colleagues about their comments as well. So disparities. The mean age of the sample has increased since 2004, and in 2014 to '16, it was 57, with a standard deviation of 16.8.

Of the ten pre-meeting comments, several felt that the disparity information was insufficient, as it only included age, gender and multi-category insurance status. And one reviewer requested the developer clarify the data that was presented. Another reviewer raised concerns about how education was categorized and used in analyses.

An important issue to me that wasn't addressed by the measure is access to the initiation and maintenance of rehab services. And it's really frankly just beyond the scope of the measure, but I do think it's important that we think about it for future measures to develop.

It would be interesting to me to know how many folks who present with back impairment are actually referred to and get coverage to receive rehabilitation services, given that we know that it makes a difference in how people do. And then that get to a rehab service that uses an assessment as sophisticated as the

photo system. So I just wanted to say, I don't want us to stop here and think that we've got the management of back pain covered.

Although this is a strong measure, I don't want us to just say, oh, we've got back pain or back impairment completely covered. Because they think there's potential for disparities in referral patterns, and particularly if we look at the relationship between the improvement and acuity at presentation. Are there disparities in that acuity of presentation that we need to understand better?

But more specific to the measure under discussion are questions about - even for those who find themselves at a rehab provider that you uses the photo system, I found myself wondering, particularly when I read the intro script, whether traditionally under-resourced populations might suffer from disparities in completing the initial online assessment, and whether there might be disparities in who's lost to follow-up.

I think some of the analyses sort of address this, but it wasn't as transparent as might be helpful to the committee. Who doesn't complete the end assessment, or the outcome - are there disparities in that? So I think again the disparities piece might be - there was a little bit more division in the comment about that part of the opportunities for improvement.

(Lisa), do you want to comment on opportunities for improvement and disparities in the…

(Lisa Suter): So this is (Lisa Suter). Thank you, I think that was a great summary. I share the concern about the disparities' information. And I know we're going to talk about feasibility later, and I know feasibility is not a must-pass criterion, but it is I think probably one of the critical things when we're talking about patient-reported outcomes.

You know, burden is such a significant issue, and there just seems to be a tremendous amount of data collection that that goes on for this measure, even though the actual patient burden is minimized by the item response theory. And so I will say they are including some aspects of social risk in the risk adjustment.

It's not clear to me -- and I may have missed, I'm sure the developer can respond -- that their accounting for response bias - which I think your point with it well taken that we should expect a response bias, and there is evidence that certain populations are less likely to respond.

And I think there's also evidence that certain populations actually utilize technology and technological surveys differently. Age already being accounted for in the model, and payer, which may get at social risk.

And that in that way we may be exacerbating disparities by not collecting information on certain populations, or collecting information on some of those populations but risk-adjusting for it and so not necessarily understanding that those populations are actually doing worse than pure populations without social risk factors, and that may actually worsen disparities over time.

(Chris): Yes, (Lisa), this is (Chris). I think I was one of the authors of one of those comments, and you said it better than I could have. I'm thinking that in future work we really - it's important to look more closely at the different disparities and what the impact might be.

(Chris): This is (Chris). Can I make a general comment on this?

Man 2: Sure, go ahead.

(Chris): It's been eating. I mean, it goes to what you're saying - and I think it was (Lisa), with a nod toward the concept of, we don't want to add additional burden. But this measure becomes infinitely more useful if either the administration or the PRO-PM is standard of care.

The patients receiving care for low back impairment and who completed the low back PRO-PM I think leaves out the most important population who may needed to have filled it out, but we have no information. And that could be very large slice.

So I'm wondering how - if a process measure exists identifying the proportion of patients who receive care for low back impairment, the proportion who get administered the survey and those who don't, because that'll really add color to what this is. Either it could be built in the measure or it's a new measure. I don't know what to do with that comment. It's just been eating me.

(Deb): Yes, (Chris), I think that's what I was referring to in my comments about sort of the wish list for future measures. I don't think that that belittles or diminishes the importance of this measure. If you are in therapy, is it leading to improvement, right? I mean it serves as an important question, right?

(Chris): Well, I guess my point is, that is solid evidence, then, to get folks to take the surveys, to be accountable for taking the survey. So I see it as a necessary QI because of the measure. Okay, I'll back off (unintelligible).

(Deb): Yes, I agree that future measurement development should look at these access issues. I think it's really important.

(Rich Anscepeli): (Chris), (Rich Anscepeli) here. (Chris) -- the other (Chris) -- I agree everything with what you said, and it loops back to what I was talking about before about the measure being fit for purpose. I think that there is enough evidence in the context of considering this measure for maintenance and re-endorse or renewed endorsement - I think you guys are making a very cogent argument for making this a standard tool in the assessment.

The notion of expanding this beyond individual clinicians into groups and into accountable care contracts, things get really scary to me, because if you've got a Medicaid accountable care organization that doesn't take into account when that patient came in and what the access to their services might be.

And then especially since the tool has the ability to differentiate beginning and end interventions we could be both increasing disparity and putting unnecessary financial risk burden on safety net providers.

So I do think we have to come back to that notion, is the measure fit for purpose? And I can't remember - somebody brilliant from the committee -- when we first had our webinar a few months ago -- said, should we use a different lens -- we meaning of the NQF -- when we're looking at measures for maintenance consideration. And I think that we are right now deep in the logic for why whoever said that was prescient.

(Chris): Yes, agreed. Especially when we refer back to evidence from 2008.

Man 2: Great. All right. Okay, just real quickly with all these comments about disparities and future measure and future work, I didn't know if the developers had any reflections that they wanted to share with us on that before we close the discussion about gap.

(Daniel): This is (Daniel). Maybe I'll just give again some clarification notes that do not address the whole discussion, just conducted, which I find really interesting. And I relate to the discussion because the whole disparity (unintelligible) and social risk structures is a huge deal that we're all struggling with and dealing with over the last years. So just for minor comments for clarification purposes.

Education was tested as one construct. It wasn't clear during our first submission. We've clarified that to SMP (unintelligible) response letter. So it was tested as you've advised us to test it. Next thing is the missing data (unintelligible), that's something that we deal with all the time, because in real life there's missing data. There's missing outcomes data, especially at discharge.

And we look at the potential bias that could come from that. And we do that in several ways. I'm not going to go over these separate methods, but it's all in the submissions. And we do that each time we run a study, each time we publish, each time we submit to NQF. We do these analyses to try to assess where the data is mostly missing - missing at random. And this was our conclusion in the case of (unintelligible).

So we know there's missing data that is a potential for bias, but we assess that using several methods, until we're comfortable that most of the missing data at discharge is missing at random.

Man 2: Okay, great. Thank you.

(Deb): Thank you, (Daniel). I mean, my question was whether - if you looked at differential loss to follow-up by education level or missing data -- both missing at all in taking the survey, and in loss to follow-up -- by education,

ethnicity - some of the factors that we know from a social drivers of health might influence that.

And I think it's sort of in there, but it just wasn't as clear that you were specifically asking about those factors and their potential influence on who gets surveyed and lost to follow-up.

And I think that's what people are asking for from the disparities' lens, is to sort of understand a little bit better specifically folks with lower health literacy as assumed by educational attainment, folks with different ethnicities, et cetera. So I'm not saying that that leads me to say this wasn't - that this is not opportunities for improvement and that you haven't taken the first steps towards disparities.

But in the next submission, I think it would be really helpful to have it a little more clearly laid out.

(Daniel): Okay, thanks for the feedback. Just a quick response. For the missing outcomes data parts, we do look at those differences, and it's part of the data we submitted. We look at the differences between those with complete or incomplete outcomes, and we did that for education as well, because we have that data to look at.

This does not address your point about access. So who gets in the first place? This addresses those that have outcomes at the end or do not. And the access is a very complex issue to analyze and get data on, but I understand your concern. So thanks for your comments.

Man 2: Okay. Great. I think we're ready to go for a vote on gap. Unless there's any disagreement? Okay.

Amy Moyer: Voting is now open on performance gaps for Measure 0425. Your options are A, high, B, moderate, C, low, and D, insufficient.

Man 3: I must admit I enjoyed the graphs you used to have on the votes. Unlike Iowa.

Man 1: We're at 16 votes. We're looking to get 19. Is anyone not able to cast your vote?

Woman 1: I had to reload. I just got it up.

Man 1: Okay, thank you. And we have (Jeri)'s vote via the chat. Thank you, (Jeri). Missing one vote. Okay, we're good.

Amy Moyer: Okay, the voting is now closed on performance gap for Measure 0425. For high, we have a count of two. Moderate we have a count of 13. Low we have a count of one. And insufficient we have a count of three. So based on these numbers, this measure passes on performance gap.

Man 2: Okay. So we need to speed up just a bit, so we are behind schedule. So let's try to keep things moving.

(Deb): So I'm just going to jump on in with reliability. So the next one - reliability. And I'm supposed to talk again about the measure specifications, which you see on the slide in front of you - the numerator, which applies -- as I mentioned before -- at three levels.

The - to get a little bit more down into the numerator specification, at the individual clinician level, it is a 12-month time period for low back pain impairment, and they must have a minimum of ten cases to be included.

And at the clinic level, it's also a 12-month time period, and there are some minimum levels also at the clinic level. The denominator exclusions are just if you're not being treated for low back impairment or if you're younger than 14.

Since last submission, the developers - and I think one of the nice things about this measure in general is that they collect data and they continue to revisit it and make modifications as needed.

They published a paper in 2018 where they updated the model to include some additional factors and variables in the risk adjustment model such as exercise history, previous treatment, medication use, some 30 specific comorbidities updated to ICD-10 codes, and post-surgical category.

They also added -- based on feedback from clinicians -- and tested the addition of three additional items to their item bank. So they had 25. Now they tested and found that it improved their IRT analysis to have the 28 items. And the items they added were difficulty using a broom, difficulty getting down to or up from the floor, and difficulty changing position quickly like sitting and standing.

So for this measure, the NQF staff preliminary rating for reliability was high. The scientific methods panel also voted reliability as high. There were three votes for high, one for moderate, and (unintelligible).

I think the one reason there was one moderate or one insufficient was just there was a question -- and this was also raised in some of the comments -- about the reliability with some of the smaller sample number of observations per clinician and then the number of occupation per clinic.

Specifically, there's sort of an accepted standard that the reliability should be 0.7 or higher. And for clinicians reporting 10 to 19 cases - although the average liability met that 0.70 standard, 42% of clinicians had reliability below 0.7. Similarly, for the 20 to 29 cases, again, average reliability was above the cut point, but still 28% of clinicians had reliability that fell below that.

So some people raised some questions about why the inclusion criteria was at ten cases as opposed to at 30 cases. And - but that was really the only question that came up surrounding reliability. Otherwise there was a lot of evidence presented that supported clear reliability. (Lisa), did you want to say anything else about that?

(Lisa Suter): The only thing I would point out is that the person who was 116 years old that was noted was probably a typo.

Man 2: Great. Any other comments about reliability?

(Don): This is (Don). I'm looking at Page 12, and I'm confused with what is on Page 12, which I think is under that the concept of reliability that we're discussing. So if you look at 16C, the boxes for yes and no are both checked for 1, 2, and 3.

(Chris): Don, let me just speak to that very briefly. These are a conglomeration of all of the methods panel evaluations. So if one panel is checked yes and the other checked no, then both boxes get checked.

(Don): Okay. All right, so that that wasn't clear. I was just confused about it. And then the other is 16D.2 is not - yes or no is not checked. Could you explain that? I'm trying to get to the test.

(Deb): It was a (unintelligible) pattern, that's why. So had to say that - it factored - if you answered basically no to 16D.1, then you were supposed to answer 16B.2.

(Don): Oh, I see. Okay.

(Deb): Does that make sense?

(Don): All right, so in other words it's N/A for this one.

(Deb): Exactly, yes.

(Don): And then just to reiterate, while the majority of people sort of agreed, there are a couple of different viewpoints -- which I hold, too -- relative to the risk adjustment, which is concerning to me on many fronts, which I won't go into.

Man 2: (Don), let's discuss risk adjustment during the validity portion of our discussion.

(Don): Okay. I'm sorry. I'm trying to figure out where we are in the document. So I apologize for that. You said reliability, and I was looking into the reliability section under testing. So I apologize - no, I'm sorry. Validity testing. I see, you're right. Okay, thank you. Sorry.

(Deb): Yes, the pink sort of blends them, so, yes. Okay.

Man 2: Okay. Should we - are we ready to vote about reliability? Okay, let's go.

Man 1: So we're going to vote on approving the scientific methods panel reliability rating, which I'll remind you is high. So if you disagree with that, then vote no. If you agree with the scientific methods panel evaluation, then vote yes.

Amy Moyer: The vote is now open for accepting the scientific methods rating of reliability for Measure 0425. Voting is now closed. We have a count of 17 for yes and two for no. So this measure passes on reliability.

Man 2: Okay, great.

(Deb): I'm going to jump in with validity (unintelligible).

Man 2: Go ahead.

(Deb): Okay. So the scientific methods panel voted validity even higher for this one. And we had four members voting it high and one moderate. There were no low or insufficient votes. The NQF staff preliminary rating for validity was also high. The developer presented data for content and construct validity, with significant improvement in the second and ninth deciles, with a distribution of residuals 58.9% to 80.9%.

And one of the nonclinical reviewers asked if that seemed clinically significant. And my answer to that is absolutely. If we could see that kind of improvement in other areas, that would be great.

There were - as we discussed already, there were comments that in the future presentations of the measure, more (unintelligible) should be done to assess potential bias resulting from ability to complete better assessment of educational levels and relationship to outcome and other social drivers of

health. They - as noted by the developers, they report that the missing data were mostly missing at random.

But I'm not someone that thinks we should risk adjust out disparities. Rather, I think it's something that really goes to - should be something that we examine and report rather than trying to include it in the risk adjuster in the measure.

Given the significant amount of data that they present that that support validity and the psychometric performance properties of the measure, I recommend accepting the preliminary rating of the NQF staff and of the SMP that validity is high.

Man 2: Great, thanks, (Deb). (Lisa), anything to add?

(Lisa Suter): No.

Man 2: Okay. And (Rich), your point's well taken about validity in different sub-populations, especially adolescents. And I think, (Don), you had a concern about risk adjustment as well. So those are noted. What other comments do we have from the rest of the community related to validity?

(Randi Oscar): This is (Randi Oscar), and I would like someone to comment on the concern or the topic that was brought up about the question - the sometimes usual hobbies, recreation, or sporting activities versus things that were more action oriented, like bending, stooping, lifting, carrying, changing position, and the effect of the actual way the question is asked on the validity.

Woman 1: Yes, so for me as a geriatrician, I think what they're talking about is whether they want - they're asking to insert some of the NAGE physical performance

items, and to do that by taking out some of the advanced activities of daily living questions.

And I think advanced activities of daily living -- particularly from a patient, both some of my more high-level functioning patients, and my younger patients -- quality of life is very much affected by your ability to do the things you enjoy, right?

So I don't know. I think - also, the developers do mention that they tested, so maybe I should let them comment. They mention that they've examined some of the NAGE physical performance items as well. So maybe the developers - do you all want to comment about those?

(Giana): Yes, this is (Giana). Is this about the questions that are more specific like sweeping with a broom, versus how are your recreational activities? Is that what this was about?

(Deb): Yes.

(Randi Oscar): Yes, they want the stooping, bending, lifting one.

(Giana): Yes, this is a great question. So first and foremost, all of the questions are scientifically good, whether they seem more general or more specific, they've all passed a really rigorous bar for being good. The background with that would be, you know, if we ask you about your recreational activities, that can mean different things.

But at the end of the day what we're assessing is, what does it mean to you, based on your recreational activities? How limited are you? As I mentioned in the introduction, though, a lot of what we do is, after we make sure everything

is scientifically good, we still want to take into account the experience of the patient, the experience of the provider.

And sometimes there are functional questions where we get a lot of feedback that, we just don't like this question, and it's causing frustration. And we get complaints from our patients. And we don't have time for that.

So sometimes we'll go in and make changes that might affect our development simply to make it a more pleasant experience for the patient and the provider. But everything that's in there is scientifically good.

(Randi Oscar): And this is (Randi Oscar) just commenting on that. And I'm thinking about the person who's working three jobs has to answer my usual hobbies, recreation, sporting activities. And they're probably thinking, are these people kidding me? And we just have to be careful that sometimes the questions that we're asking are not reflective of all populations.

And my second follow-up to that would be, when I think about, say, think back issues, or neck issues, or any kind of movement issues, I would love see standardization. So percent moving, right? Your ability to - how flexible were you before versus now? The consistency in thinking about this measure in terms of other ones, I think, is a hat that I would recommend we wear.

(Daniel): This is (Daniel). Can I give a short comment?

Man 2: Great, yes. Please keep it short if you can.

(Daniel): Yes, sure. So just to reinforce what (Giana) said -- and thanks for these questions -- all items are checked not only for their structural validity -- we're talking about validity here -- which means they all relate to one overall main

construct. So they relate well to one another and to the construct we want to measure. We also look at differential item function, and that's also in the submission.

So we want to make sure that people - patients from different groups perceive the difficulty level of the item and understand the item in a similar manner. So that's one of the validity testing one of the validity testing presented.

(Tracy Wong): Hi, this is (Tracy Wong). I was wondering if someone could comment on this. How do you account for the weight line between time points? So I see that BMI is used as a risk adjuster, but it's unclear to me whether that's only at the beginning or whether you're also counting for weight loss between time point A and B. I think in particular about a cohort of women who have just had a child and are naturally losing weight, and just how you thought about that.

(Daniel): Yes, thanks for this question. We do not account for that. BMI is a comorbidity. It's a preexisting condition like all other comorbidities that we assess. We need to have all of those at the beginning of care in order to adjust for them. So it's a great point. But no, we do not have that data, and therefore we do not test it.

(Giana): Okay. If I could add - that is a great point. One of the things that we're looking at adding in the future is assessing whether or not the patient is pregnant, because that -- besides weight loss -- can have other factors as well.

Keep in mind -- it is a good point -- but with physical and occupational therapy episodes of care, they tend to be a lot shorter than, say, a medical episode of care. So weight loss, but it's possibly less likely in that shorter time.

(Don): So this is (Don). I'll just chime in now as a patient, given that I've had three spinal surgeries in the past six years now. And the one thing that I'm struggling with here too is that there isn't any sort of cross-correlation with the ability and interests and consistency and sustainability of an individual patient following a very specific set of exercises related to this prescription. Because most of the action with physical therapy doesn't occur in the physical therapy session, right? It occurs outside of the physical therapy session.

And I just know from a patient care standpoint, too, that the amount of consistency is all over the map in terms of how patients take the things that they're supposed to do seriously and do them. So I'll just leave it at that.

(Giana): Would you like a developer response to that, or…

Man 1: So we really need to move forward. We're about ten minutes over on this vote, and we're leaving no time for the next one.

Man 2: Yes, sorry. So if there's any other burning questions on validity - I mean, great discussion but it's not really central to our voting. So…

(Deb): And again I want to point out that the preliminary rating of the NQF staff and the SMP for validity was high.

Man 2: Right, were both high.

(Deb): Based on the 140-page packet.

Man 2: Okay. Great. Let's vote on validity.

Man 1: So we're going to do separate vote on this. I feel like the discussion has been sufficiently robust that we vote the high, moderate, low or insufficient on validity. So the vote is now open.

Man 2: Sure.

Amy Moyer: Okay. Voting is now closed. We have seven for high, ten for moderate, one for low, and one for insufficient. Based on this, the measure passes for validity.

Man 2: Okay.

(Deb): So I'm going to jump into to feasibility. Sorry I'm rushing (unintelligible)…

Man 2: No, please rush.

(Deb): So feasibility. So the photo system uses computer-assisted testing format and IRT mapping of the items to decrease respondent burden. And the developer provides data that the average time to complete is five minutes. So the only potential limitation that was raised in the comments was the proprietary nature of the survey.

However, there is free access to the components needed to calculate a reportable score that can be found on the photos Web site, and I actually went to it and checked it, and it is there. And you can purchase or pay for additional support, such as assistance with data collection, scoring, and report generation for a monthly fee that's not outrageous. It's $20 a month per clinic or $15 a month for a provider.

And there's also a higher level of support that's available, but there is baseline free access to the components needed to calculate the score. Another comment raised a question about proxy completion differences, but again, the current performance metrics of this measure are really high, and it includes proxies.

So I'm not - although it might be interesting to see it pulled out, I'm not as concerned about it affecting the feasibility of using the measure. Overall, I therefore support the NQF staff's preliminary rating for feasibility at high.

Man 2: Great, thanks, (Lisa) and (Lisa)?

(Lisa Maurice): This is (Lisa Maurice). I don't have anything to add, except thanks, (Deb), great job.

Man 2: Yes. Great. Comments about feasibility from the committee? Okay. Are we ready to vote on the feasibility, then I believe?

Amy Moyer: The voting is now open on feasibility for Measure 0425. Your options are high, moderate, low and insufficient. The voting is now closed. We have 11 for high, six for moderate. low has one, and insufficient has zero. Therefore this matter passes on reliability.

(Deb): Feasibility.

Amy Moyer: Feasibility. Correct, sorry.

(Deb): All right, let me jump into use, next category. So I'll start by saying that the NQF preliminary rating was pass for use. And basically the measure is currently publicly reported and used in accountability programs. It's used in CMS's PQRS, the MIPS program, the physical therapy provider network,

therapy partners use this photo outcomes and value-based contracts with payers.

The providers can actually get real time reports on their individual patients, and they seem to be providing the developers with a lot of feedback on desired improvement. So I think that is an indication that people are paying attention to it and using it. This was where I was going to talk about the (unintelligible) performance items that we've already discussed that issue, as an example of feedback coming back to the developers.

The pre needing comments about use were mixed, with one reviewer commenting, "it needs field testing for the impact of comparative non-risk adjusted results to determine discriminatory accuracy" And another commented that "Attributions to clinician or clinic appears weak." However, others noted it important for goal setting and treatment planning.

And you know it does seem from a practical clinical perspective that if you prescribe therapy, it's with the expectation that that is a therapy, the more improvement people are going to get in the condition associated function.

Knowing the extent to which that's occurring -- and it varies across providers for one reason -- seems a very relevant use. And motivating patients and helping them here is part of effective therapy, in my framework, having seen several different therapists work with my adolescent child.

So I would, again, in summary, support the NQF preliminary rating of pass for use. Should I go on to usability before we have more discussion? Or do we want to stop (unintelligible)…

Man 2: Yes, that's fine. Go ahead.

(Deb): Okay, usability. Data analyses show improvement over time in provider and clinic performances on the metric. Small improvement, but improvement, suggesting that the data may be used for improvement activity. CMs is planning more evaluation based on 2019 data.

It was noted by a couple of commenters as presented by the developer that one of the diagnostic categories that has been historically used in risk adjustment was used to deny a military application. So that being considered - the developing that they're thinking about not using that category moving forward.

Another reviewer raised concerns about attribution again. Agree with comment that one reviewer made that the benefits outweigh the risks, and that the measure can be used to improve monitoring of outcomes. The NQF staff preliminary rating for usability is moderate. So (Lisa), any…

Man 2: Thank you, (Deb).

(Lisa Suter): I don't have anything to add.

Man 2: Great.

(Deb): Other (Lisa)?

Man 2: Okay, comments from the committee on usability and use? Okay, let's vote.

Man 1: So we'll vote on use first. Voting is now open on Measure 0425 for use.

Amy Moyer: And your options are pass and no pass. The voting is now closed. We have 16 for pass and three for no pass. Based on this, the measure passes on use.

Man 1: Let's proceed directly to the vote on usability. So voting is now open for usability. Your options are high, moderate, low or insufficient.

Amy Moyer: Voting is now closed. We have three for high, 14 for moderate, zero for low, and one for insufficient. And based on that, this measure passes usability.

Man 1: Okay.

Man 2: Okay (unintelligible).

(Deb): One category left. And that's related or competing measures. And as noted in the materials, the measure is related to but not competing with seven other measurements that look at functional status change for patients with other types of musculoskeletal impairment, including one category with general orthopedic impairment. So again it's related to but not competing.

Man 2: Okay, any other discussion about that? Okay.

Man 1: All right, let's vote on overall endorsement.

Man 2: All right.

Amy Moyer: The voting is now open. Your options are yes or no. Voting is now closed. We have 17 yes and two no. Therefore, the measure passes on overall.

Man 2: All right, very good. Well, we only have a limited amount of time, but let's get as far as we can on the other measure. So pivoting directly over to our

measure developer for Measure 0291, Emergency Transfer Measure. (Jo Klinger), area you on the line?

(Jo Klinger): Can you just comment, if we don't finish, what the plan is (unintelligible) clock stops?

Man 2: Yes, we'll have to reschedule an additional call.

(Jo Klinger): All right. Does it make sense to just do the additional call? Because there's twelve minutes.

Man 2: Well, let's get as far as we can.

(Jo Klinger): All right.

(Ira Marcovitz): Can I offer a comment? This is (Ira Marcovitz) the other co-developer. I really agree with the last comment. I don't think it makes sense to do this for five or ten minutes. The last call took an hour and 20 minutes.

I really do think if we're going to have to have a call, we should just do it in one whole chunk, and that'll help us a lot understand your concerns and give us appropriate time to respond to your comments. That's just my suggestion.

Man 2: Okay, I'll defer to the cochairs on this one.

(Chris): Yes, this is (Chris). In the past, anything that we can do ahead of time - we will have to recap a little bit, but if we're running low on time in the next call, I think taking advantage of the ten minutes now might be helpful just to at least start the thinking. And I'm sorry that people have to be on for more than one call.

(Jeri Lam): Hi, this is (Jeri). I think there are a lot of issues in this measure, and this is one of the very few care coordination measures. I would wait and let's have a total discussion. Plus we I think need to have public comments in less than two minutes or three minutes, so I'd rather do it as a whole.

Man 2: Okay. Well, you're running it, (Jeri), so it's your call.

(Jeri Lam): Yes. And I think we had a robust discussion, and let's hope we can do it in an hour this time.

Man 1: Okay. Let's go ahead and open it up for public comments, then. The lines are all open for any members of the public or the NQF membership to make any comments related to Measure 0425.

If you wish to enter your comment as a chat, you may do so through that modality as well, and NQF staff will read your comment. All right, seeing none, let's go ahead and adjourn for now. And we'll reconvene to discuss Measure 0291 as soon as we can get some dates on the calendar.

Thanks, everyone, for your participation, (Chris), for leading (unintelligible) and thanks to all of your for your careful review and consideration.

((Group)): Thank you.

END