NATIONAL QUALITY FORUM

+ + + + +

BEHAVIORAL HEALTH PHASE 3
STANDING COMMITTEE MEETING

+ + + + +

WEDNESDAY
OCTOBER 1, 2014

+ + + + +

The Standing Committee met at the National Quality Forum, 9th Floor Conference Room, 1030 15th Street, N.W., Washington, D.C., at 8:30 a.m., Peter Briss and Harold Pincus, Co-Chairs, presiding.

PRESENT:
PETER BRISS, MD, MPH, Medical Director, CDC, National Center for Chronic Disease Prevention and Health Promotion
HAROLD PINCUS, MD, Director of Quality and Outcomes Research, New York-Presbyterian Hospital, The University Hospital of Columbia and Cornell
ROBERT ATKINS, M.D., MPH, Senior Medical Director, Aetna Medicaid
CAROLINE CARNEY DOEBBELING, M.D., Msc, Chief Medical Officer, MDwise, Inc. *
MADY CHALK, PhD, MSW, Director, Policy Center, Treatment Research Institute
DAVID EINZIG, MD, Medical Director of Child Psychiatry, Children's Hospital And Clinics Of Minnesota
JULIE GOLDSTEIN GRUMET, PhD, Director of Prevention and Practice, Education Development Center/Suicide Prevention Resource Center/National Action Alliance for Suicide Prevention

CONSTANCE HORGAN, Sc.D., Professor and Director, Institute for Behavioral Health, The Heller School for Social Policy and Management, Brandeis University
LISA JENSEN, DNP, APRN, Associate Director Workforce & Leadership, Office of Nursing Services, Veteran's Health Administration
DOLORES (DODI) KELLEHER, MS, DMH, Principal, D Kelleher Consulting
KRAIG KNUDSEN, PhD, Chief, Bureau of Research and Evaluation, Ohio Department of Mental Health and Addiction Services
MICHAEL LARDIERI, LCSW, Assistant Vice President Strategic Program Development, North Shore-LIJ Department of Psychiatry
TAMI MARK, PhD, MBA, Vice President, Truven Health Analytics
RAQUEL MAZON JEFFERS, MPH, MIA, Director of Health Integration, The Nicholson Foundation
BERNADETTE MELNYK, PhD, RN, CPNP/PMHNP, FAANP, FNAP, FAAN, Associate Vice President for Health Promotion, University Chief Wellness Officer, Dean and Professor, College of Nursing, Professor of Pediatrics & Psychiatry, College of Medicine, The Ohio State University
LAURENCE MILLER, MD, Senior Psychiatrist, Arkansas Medicaid, Arkansas Medicaid
DAVID PATING, MD, Chief, Addiction Medicine, Kaiser Permanente
VANITA PINDOLIA, Pharm.D., VP, Ambulatory Clinical Pharmacy Programs, Henry Ford Health System/Health Alliance Plan
RHONDA ROBINSON BEALE, Medical Physician, Former Chief Medical Office at Optum now Health Care Consultant, Health Care Consultant
HENA SIDDIQI, M.D., Medical Director, Broadlawn Manor Nursing and Rehabilitation

LISA SHEA, M.D., D.F.A.P.A., Deputy Medical Director, Quality and Regulation, Butler Hospital
JEFFERY SUSMAN, M.D., Dean, Northeast Ohio Medical University, Northeast Ohio Medical University
MICHAEL TRANGLE, MD, Associate Medical Director for Behavioral Health, HealthPartners
BONNIE ZIMA, MD, MPH, Professor in Residence, Child and Adolescent Psychiatry, UCLA Semel Institute for Neuroscience and Human Behavior
LESLIE ZUN, MD, MBA, Chair, Department of Emergency Medicine, Mount Sinai Hospital

NQF STAFF:

POONAM BAL, Project Analyst
HELEN BURSTIN, MD, MPH, FACP, Chief Scientific Officer
LAURALEI DORIAN, Project Manager
ANGELA FRANKLIN, Senior Director

ALSO PRESENT:

SAM TIERNEY, AMA PCPI
KENDRA HANLEY, AMA PCPI
SARAH HUDSON SCHOLLE, NCQA
MEREDITH JONES, AMA PCPI
TONI KAY, AMA PCPI *
JUNQING LIU, NCQA
MICHAEL MURPHY, Massachusetts General Hospital *
KAREN PIERCE, AMA PCPI *
BOB REHM, NCQA
SARAH SAMPSEL, Azul Quality Solutions, Consultant to NQF

* present by teleconference

C O N T E N T S

P R O C E E D I N G S

(8:35 a.m.)

Introduction and Disclosure of Interest

MS. DORIAN: Good morning, everyone. Welcome to the Behavioral Health in-person meeting. This is Lauralei Dorian, and it's good to finally see so many of you here in person, put faces to names. Many of you we've been working with over previous years, many are new to NQF. So we welcome you all on behalf of all of us here at NQF.

Before we get started I just wanted to do a few logistic -- note a few logistic things. The restrooms are located outside of these doors to the right-hand side. People will be able to direct you if you have any questions. There should be somebody out there at the desk at all times if you had any questions about your reservations, or booking a taxi, or anything else.

A reminder that all of the lines are open to the public, and there will be time for

public comments at dedicated times throughout the meeting. All materials are available on SharePoint so if you have your laptops with you you'll be able to access them on SharePoint. If you have any problems please come tap one of us and we can help you out.

We have used a method in the past that has seemed to work to -- that we ask that you put your name tag on its side if you have comments, so that will allow the co-chairs to identify you and make sure that everybody's able to comment in time. For those of you who are on the phone, you can follow along on the public Webinar, and ask questions or comment using the chat box feature.

We have made dinner reservations tonight at 6:30 -- or actually I think 6:00 p.m. at Mio, which is just around the corner here. It's a contemporary Latin American cuisine, so I know a lot of you indicated that you'd like to go. If anybody didn't let me know there's still openings, so come and find

me during lunch.

We will have breaks if you have -- you should have the agenda in front of you. We will have breaks at 10:45, you know, give or take, 1:10 lunch and 3:15 another fifteen minute break. So we'll really try to stick pretty closely to the agenda, because we do have a fair number of measures to get through. And Howard and Peter have been with us for a long time and they're experts at keeping us all on schedule, so feel confident. So I've introduced myself and then I'll have the rest of the NQF team introduce theirselves.

MS. FRANKLIN: Thank you so much, Lauralei.

I'm Angela Franklin, Senior Director for the project, and I'll let Poonam introduce herself.

MS. BAL: And I'm Poonam Bal, I am the project analyst on this project.

MEMBER SAMPSEL: And I am Sarah Sampsel, I'm a consultant to NQF on this

project.

MS. DORIAN: Great. And now we would love to hear from all of you. I'll turn it over to Harold first and then we'll go through our roster and get to hear a little bit about your backgrounds and where you're coming from.

CO-CHAIR PINCUS: I'm Harold Pincus, I'm a professor and Vice-Chair of psychiatry at Columbia University and also Director of Quality and Outcomes Research at New York Presbyterian Hospital. I'm also a senior scientist at the RAND Corporation.

And I'm delighted to welcome everybody here. We have a lot of work to do, there is a whole long list of measures. And you know, I think at some point you guys will sort of make the point that NQF has gone through a fair amount of redesign of its operations, especially of the endorsement process, and it's becoming a bit more formal in a number of ways. And I think in that way

more rigorous and more valid, I think, in terms of the process, and certainly more standardized.

And so we're going to be sort of going through this in a somewhat new way, although a lot of the same themes come through. The criteria have not changed very much and really we're going to be thinking hard about each of the criteria as we go through each of the measures. That's probably the most important, I think, distinction in terms of how things have shifted a bit, so we really do get a clear thoughtful approach at each of the measures and to think about the extent to which the measure meets or does not meet the specific criteria.

And then also we'll be voting on each of the criteria as we go through this. So just to kind of introduce that process, and just be aware as we go through, we'll get into more detail as we go through the first few and people get a better sense of, you know, what

we really mean.

MEMBER SUSMAN: You're off-mic a little bit.

CO-CHAIR PINCUS: Oh, sure. I just moved it closer. Sorry. It's -- yeah, and I got in about 1:00 o'clock last night and sort of -- after a day of being on conference calls so I have no voice. Okay. Well, thanks. And so why don't we continue to go around. Bob?

MS. DORIAN: And if everybody could, yeah, use your mics, because it's being recorded.

MEMBER ATKINS: Hello, my name is Bob Atkins. I am a psychiatrist, Senior Medical Director with Aetna Medicaid, and the behavior health lead for Aetna Medicaid nationally.

MEMBER TRANGLE: Hi, I'm Michael Trangle, psychiatrist and Medical Director for an integrated system in Minneapolis-St. Paul called Health Partners.

MEMBER HORGAN: Hello, I'm Connie Horgan, I'm a professor at the Heller School

of Brandeis University and Director of the Institute for Behavioral Health. I'm a health services researcher and do a lot of work in how to improve the system so that measures can be adequately implemented. Thank you.

MEMBER SUSMAN: I'm Jeff Susman, I'm the Dean of the College of Medicine at Northeast Ohio Medical University which is a small community-based medical school and actually at Rootstown in Cleveland, and we cover all of Northeast Ohio. My connection here is when I did do research and had some interest in other things than shuffling papers, I was interested in implementation research around depression and mood disorders.

MEMBER SHEA: Good morning. I'm Lisa Shea, I'm the Medical Director at Butler Hospital, a freestanding psychiatric hospital in Providence, Rhode Island, and also a clinical associate professor at the Alpert Brown Medical School.

MEMBER SIDDIQI: Hello, I'm Hena

Siddiqi, Geriatrician, Medical Director at Rudloe Manor which is also onsite with a tertiary psychiatric hospital, South Oaks Hospital. It's part of the North JRLA system.

MEMBER KNUDSEN: Hi, I'm Kraig Knudsen, I am the Chief of the Bureau of Research and Evaluation at the Ohio Department of Mental Health and Addiction Services.

MEMBER LARDIERI: I'm Mike Lardieri and I'm AVP of Strategic Program Development at the Northshore LHI Health System in New York. Previously I was with the National Council for Behavioral Health, my last time here, and I do a lot of work with HIT and integration of behavioral health and physical health.

MEMBER MELNYK: Good morning. I'm Bernadette Melnyk, I'm from the Ohio State University. I am the university's Chief Wellness Officer and Dean of the College of Nursing. I'm both a pediatric nurse practitioner and a psychiatric nurse

practitioner.

MEMBER ZIMA: Hi, and I'm Bonnie Zima and I'm a professor, UCLA, and I'm a child psychiatrist and health services researcher.

MEMBER PINDOLIA: Hi, I'm Vanita Pindolia. I am the Vice President of the Ambulatory Clinical Pharmacy programs for Henry Ford Health System and Health Alliance Plan. And I think my relationship with this committee has been, with the work I've been doing right now with the Medicaid plan to develop their dual eligible clinical programs but also on the provider side and the health plan side to improve their quality metrics for both related to behavioral health and others.

MEMBER KELLEHER: I'm Dodi Kelleher, I'm a behavioral health clinician and independent consultant. I work primarily with large, self-funded employer health plans, helping them integrate behavioral health into their medical and disability benefits.

MEMBER ROBINSON BEALE: Hi, I'm Rhonda Robinson Beale, I'm a healthcare consultant working currently with health plan in Illinois with the dual eligibles helping them to rule out and design and roll out their care management process for not only medical but also behavioral health.

MEMBER JENSEN: Good morning, I'm Lisa Jensen, I'm a psychiatric advance practice nurse. I work for Veterans Health Administration, Office of Nursing Services here in D.C., but I'm a virtual employee and I work from my home in Salt Lake City.

MEMBER EINZIG: Hi, my name is David Einzig, I'm a child psychiatrist at Children's Hospitals and Clinics of Minnesota, which is a large, freestanding children's hospital. I'll be Medical Director this coming year. My background, I did the combined training program in pediatrics and psychiatry so I have a dual role. One of my passions has to do with collaborative care models integrating

psychiatry and behavioral health into primary care clinics and specialty clinics, and so I'm in the pediatric clinic one day a week also.

MEMBER MILLER: I'm Larry Miller, I live in Little Rock, Arkansas, I'm a psychiatrist. I'm Senior Psychiatrist at a division of medical services at DHS, which is Medicaid, and I'm also clinical professor of psychiatry at the University of Arkansas for Medical Sciences.

MS. DORIAN: Thank you. And do we have Caroline, or any other committee members on the phone?

MEMBER DOEBBELING: Hi, this is Caroline.

MS. DORIAN: Hi, Caroline.

MEMBER DOEBBELING: Good morning. I am Caroline Carney and I am the Chief Medical Officer of Medwise, Inc., a health plan serving Medicaid and marketplace populations. I am an internist and psychiatrist and a researcher by training and I'm happy to be

here.

MS. DORIAN: Thank you. And anyone else on the phone from the committee? Okay. Nice to see we have a few new committee members. Welcome, we're just getting started, just in time. If you wanted to -- that's Tammy. If you wanted to introduce yourself briefly by just pushing the speak button on your microphone that would be great.

MEMBER MARK: Hi, I'm Tami Mark from Truven Health Analytics.

MS. DORIAN: Great. Thank you. And Raquel? Welcome.

MEMBER MAZON JEFFERS: Hi, I'm Raquel Mazon Jeffers from the Nicholson Foundation.

Project Introduction and Overview of Evaluation Process

MS. DORIAN: Great. Thank you. It's good to see such a wonderful turnout. It apparently is a little bit difficult for the people over the phone to hear, so I have to

remind myself of this as well, but if you could just make sure to bend the microphone down and speak close to it, that would be very helpful.

So if we could go to the next slide, we just wanted to go over some ground rules for today's meeting. During the discussion committee members hopefully are prepared and you've reviewed all of the measures beforehand. You have measures assigned to your specific workgroup, and you were designated a lead or secondary discussant. But by this point we do expect that you have -- you will have reviewed all of the measures that were submitted to this phase of work.

We'll ask that you base your evaluation and recommendations on the measure evaluation criteria and guidance. You should have in your packet in front of you a brief overview of that guidance, and we'll be sure to go through it and walk you through to make sure you understand this, particularly for the

first measure, as we go through throughout the day.

We'd like you to remain engaged in the discussion, this is an open forum. Feel free to ask anything of each other and the developers. We're fortunate to have the developers here and they will be up here seated at the table or over the phone when their measure is being discussed.

You'll have about two to three minutes to introduce their measures so at the beginning of each measure they'll give the brief introduction and then we'll turn it over to the lead discussant to summarize what the workgroup had discussed and any questions or concerns that were raised on the workgroup call. And those summaries are found on SharePoint in those measure documents. So it's those same documents that you've been looking at, we've just updated them. They read the measure number and then ALL in capital letters.

So, as is the case with committee members, developers can put their cards up if they have any questions or comments to -- or if they want to respond to the committee.

And during -- it is important to note that during these evaluations, almost all the time the committee members make suggestions about how the measures might be improved in the future, and we do encourage that certainly. But we want to remind you that when you go to vote, you are voting on the measure as it is specified currently. So I'll turn it over to Poonam to discuss more of the role of the standing committee.

MS. BAL: So I have a lot of computers in front of me, so it's going to be a little difficult. But basically the role of the standing committee is to act as a proxy for NGS multi-stakeholder membership group. You all bring a different aspect to this -- still not loud enough? Sorry.

You all bring a different knowledge

base and expert level to this committee, and that's why you were brought together. You will be either serving a two or three-year term and we will determine that tomorrow by random selection.

And basically the role is to work with NQF staff to achieve the goals of the project, which is mainly to review all the measures and evaluate them against the criteria. And, basically, determine if the criteria is met and rationale behind determining them, if they're met.

You'll be making the recommendations to the NQF membership and moving forward will respond to comments from both the public and membership, and also responding to any feedback from CSAC. So we'll go through -- I think we've gone through the timeline before and so there are a lot of different procedures. So throughout the process, it's just not at this point, once we get comments in, once we go through CSAC, so on, you'll be

asked to give feedback on those results. And then, overall, just oversee the portfolio of the behavioral health measures.

Okay. And so we do have a new function of the standing committee. This may not be so new for behavioral health since we are in our third phase now, but it's a new concept for NQF where the committee will continue to -- we won't seat a new committee every time we get work for the topic area, it will be standing. So if anything comes up, if we do another phase, or we need to do a temporary expert panel, the standing committee would be used for that purpose.

Basically, the same responsibilities that fall under what we just spoke about, but overseeing the portfolio is the main goal, we want to continue this process. So we want to make sure that all the measures are viewed together instead of just at one time. And so knowing what measures are included in the portfolio, understanding their importance, and

we'll briefly go over what measures are currently in the portfolio in a little bit. Understanding the issues that come with standardization, harmonization, identifying measurement gaps in the portfolio, seeing what measures we need. And then just being aware of what's going on in that topic area where you really come in, being the experts.

And then just also being open to external input on that portfolio, so if you get any feedback on these measures or if you learn about the use of these measures, bringing that forward is also important. Just overall, just you are now the keepers of this portfolio, so just keeping it up to date and maintaining the harmonization.

MS. DORIAN: Great. And now Angela will go over a brief portfolio review.

MS. FRANKLIN: Thanks, Lauralei.

And as Lauralei mentioned, this project has been underway for at least -- we are in our third phase now. And we started

out with Phase 2 in 2012, and during that time we endorsed ten measures, and you can see those before you in the subject areas of tobacco, alcohol, substance use, adherence to medications, health screening and assessments, post-care follow-up following hospitalization. And the key part of the behavioral health portfolio is that we're looking for measures covering all of these topics including tobacco use, alcohol use, substance use, as well as behavioral issues such as ADHD, hyperactivity disorder, as well as measures that have to do with the screening for people with serious mental illness.

So in our first phase you can see the measures here that we have endorsed in these areas. And I'll move on to the next slide. In our second phase we had a lot of carryover in the tobacco and alcohol portions of the portfolio, and ended up endorsing 20 measures in those two areas, tobacco, alcohol and substance use.

Next slide. And we did move on to inpatient psychiatric services, which we call the hospital HBIPS, and looked into also the areas of depression and major depressive disorder in Phase 2 of this project. We had measures that had to do with screening on admission for hospital-based inpatient psychiatric services all the way through the process to discharge, and post-discharge continuing care, and follow-up for depression.

No, we're good. And in this phase we're looking again at tobacco use, alcohol use, substance use, as well as ADHD. And moving on to the next slide, depression and major depressive disorder, and health screenings and assessments for people with serious mental illness, which is the topic we touched on in Phase 1 of our project here. So we're coming full circle.

One of the things that we'll be asking you to do, as you review these measures is keep in mind the measures that we already

have in the portfolio, as well as identify gaps that you see in the portfolio, in terms of what kinds of measures we need, in terms of process measure or outcome measures, as well as measures in areas where we have definitive gaps.

Next slide.

CO-CHAIR PINCUS: Just one question, Angela?

MS. FRANKLIN: Yes.

CO-CHAIR PINCUS: What's the measure that you just went through? Does that include all of the measures that are in the portfolio of endorsed measures, or just the ones that have been in the three phases?

MS. FRANKLIN: These are just the ones that have been in the three phrases at this point, yes.

CO-CHAIR PINCUS: Yes, it would be useful at some point during the meeting if we could sort of see the list of all the measures that are in the endorsed portfolio.

MS. FRANKLIN: And we'll definitely tee that up, at Day 2 we're going to have a in-depth discussion of the gaps and we'll tee that up for you so that you can see the entire portfolio.

Questions about the portfolio or our gaps discussion that's to come on day two? Okay. So then that moves us on to our evaluation of the measures. And before we get started, I wanted to tell everyone about our new disclosure process. We gave to each of you a measure-specific disclosure of interests form, which everyone filled out, thank you for your cooperation in doing that. And we have before you on the table the results of the completion of that disclosure of interest, with regard to specific measures.

And please note that we have conflicts listed here. Harold Pincus and Constance Horgan have conflicts with the NCQA measures, and then Michael Trangle has a conflict with the Minnesota Community

Measurement measures. And Bonnie Zima has a conflict with the Mass General Measure, 0722 Pediatric Symptom Checklist.

And just to give you a little background, NQF is in the process of reviewing and revising their comprehensive disclosure of interest policy, that is being rolled out in the coming months. Generally when we get together for these in-person meetings, we have our general counsel go around the table and ask everyone to disclose any potential interests. And that policy is being revised we won't be doing it at this particular sit-down, we'll be doing it in January, at our post-commenting call.

And for this particular meeting we just felt it was most important to identify the measure-specific conflicts at this time. And we've identified those, those are on the table before you, if you need to refer to those. Also if at any time you feel like you have a question about potential conflicts or

disclosures that need to be made, feel free to come to any NQF staff and make those at any time. Are there any questions about the disclosures of interest before us?

Okay. So with that I'll move on into our evaluation of measures, and just the process for today. We'll have, as Lauralei mentioned earlier, as we tee up each of the measures we'll start with a brief introduction by the developers. They're behind me at this time, but when their measure comes up they'll be coming to the table here so that we can all have a good conversation with them at the table. Assigned discussants are going to speak to the measures first, criterion by criterion, then throw it open to the workgroup members if they have additional comments, or the secondary discussant if they have additional comments. And then we'll open the floor for the full committee discussion.

And again, we had asked that each committee member review each of the measures,

so everyone should feel free to jump in and discuss the measures, unless they have a conflict. So then we will vote on the criterion and we'll proceed through each criterion in that fashion. Are there any questions about this process? Yes? Please use your mic.

MEMBER KELLEHER: Are we going to get a refresher on how to use the voting mechanism here?

MS. FRANKLIN: Yes, we will. Before our first vote we'll definitely do that. And I'd also like to tell everyone it usually takes us a little bit longer to get through the first measure so don't be nervous as we go through and we remind everyone of the criterion as we go through, and the voting process. Harold?

CO-CHAIR PINCUS: Just one question is, and we discussed this before, but at what point do people bring in and refer to questions and issues that came up during the

workgroup meetings?

MS. FRANKLIN: Right at the beginning, in the introduction of the measure by the lead and secondary discussants. So Harold was asking at what point do we bring in the discussion points that were made at the workgroup level, and I was saying right at the beginning, when you introduce a measure, tell a little bit about what it is and thoughts of the workgroup.

MS. DORIAN: And as I had mentioned before, because some of the workgroups were quite a long time ago at this point, so to access those summaries you can click on the measure number and then each section is broken down by criterion, so importance. You'll see all the survey results and then you'll see the workgroup comments. And is everyone able to access the internet and the SharePoint sites? Does anyone have any -- yes?

MEMBER MAZON JEFFERS: I just need the password for the WiFi.

MS. BAL: It's NQF, all in caps, guest. This is in all lowercase, sorry.

MS. DORIAN: Just going through SharePoint, if you want it to. I mean, if you recall the -- we'll also bring all of these documents up on the screen as we work through them. So if you can't access them that's perfectly fine.

MS. BAL: And just a reminder, if you have any questions or comments, please click the speak button so we can get it in the transcript. Thank you.

MEMBER SUSMAN: Are we going to have an opportunity tomorrow during the more general discussion to talk about the pros and cons of specific measures and disease states, versus broad measures?

MS. DORIAN: Yes, that's definitely --

MEMBER SUSMAN: Because it seems to plague us over and over again, and certainly --

MS. DORIAN: That's true. It's not unique to this committee but that's certainly something that we would want to discuss in the gaps discussion, and the idea of what sorts of measures we would like to see coming forward in the future.

MEMBER SUSMAN: Thank you.

MS. FRANKLIN: Michael?

MEMBER TRANGLE: This is a comment I made during one of the phone calls, too. But I don't see it on the agenda, but from the vantage point of someone in practice and kind of running a system, we have to look at how much clinicians are actually doing, you know, in the office. I think it would be wise for us to somehow, at some point, step up to 10,000 feet or 50,000 feet and say what's the overall impact if we did all of these measures?

You know, sort of like the IRS, that's how much time it takes to do this document, you know? And begin to sort of

think about how we don't replicate what happens in primary care where if they did all the required measures if you were working 18 hours a day, you know? And I don't know the answer to that but I think it should be in our mindset and some of our discussions.

MS. DORIAN: Definitely. Thank you. It's good to remember that, as sort of a framework, moving forward. Did David have a question? No? Okay.

MS. FRANKLIN: Did you want to give us a quick introduction to yourself?

MEMBER PATING: Yes. I apologize for being late, my letter had 8:30 as the start time. I'm David Pating from Northern California Kaiser Permanente. I'm Commissioner of Mental Health for the State of California, Commissioner of Health for San Francisco, and I oversee evaluation for 1.3 billion in California State investment in mental health, a Mental Health Proposition 63.

MS. FRANKLIN: Thanks, David. And I

just had a note about the agenda. Before we get started, actually, what we'd like to do is start with the Pediatric Symptom Checklist Measure first. We're awaiting our second Co-Chair, who will be able to chair the first measure. So we are going to start with the Pediatric Symptom Checklist Measure first. And apologies. Could we have the developer for that speak to it, if they're in the room or on the phone?

CO-CHAIR PINCUS: And just to that, Tami, you're going to be the lead discussant for this one, and Michael, you're going to be secondary.

Child and Adolescent Measures

#0722: Pediatric Symptom Checklist (PSC) and Psychosocial Functioning

MR. MURPHY: Hi, it's Michael Murphy. Can you hear me okay?

MS. DORIAN: So we can hear you, Michael.

MR. MURPHY: Great. So first of

all, I thank you for inviting me to be here on the call, and for all the work you've done looking at the PSC. We're very grateful that the PSC is a measure that NQF has endorsed, and I think today I'm just supposed to introduce it a little bit. If you won't mind, I'll just read from the first paragraph of the first document to start.

The PSC is a brief parent report questionnaire that's used to assess overall psychosocial functioning in children three to sixteen -- eighteen years of age. It was originally developed to be a screen that would allow individual pediatricians to identify individual patients who had problems in their caseload. And due to its widespread use in large systems, it's increasingly been used as a quality measure and as an outcome measure.

There's been over 150 studies of the PSC over the past three decades and in the last year or two the research has really ramped up, in terms of large-scale systems

like a big statewide program in Massachusetts, and a national program in Chile.

Just a couple more things, it's both a process measure and an outcome measure. As far as we know, it's one of just a handful of child psychiatry measures that's been endorsed by either NQF or CHIPRA. And it's also one of the few measures that bridges pediatrics and mental health, so it's about mental health in a pediatric setting.

So that's pretty much all I had to say. We have a one-page summary that Lauralei asked us to prepare summarizing the recent evidence which is quite strong, we think. And I don't know if, Lauralei, you were able to get that to committee members.

MS. DORIAN: It is posted to the SharePoint page. But actually if you could sort of summarize that in just a few sentences, that would be great.

MR. MURPHY: Yes. So the State of Massachusetts mandated routine psychosocial

screening as a part of EPSDT, people really had to do that but they had a great data system that allowed tracking of what happened. This is going back seven or eight years now. In that time approximately two million kids have been screened and Karen Hacker and her colleagues at Cambridge Health Alliance got a hold of the State Medicaid datasets and tracked what happened. And a couple of papers published in pediatrics show that the screenings did identify about 30 or 40 percent of the positive screens had not been receiving services, and about 30 percent of the kids that were newly positively screened went on to receive services. So we have confirmation in some large datasets of screening that seems to be associated with better outcomes.

I know time's limited so I'll stop talking.

CO-CHAIR PINCUS: Okay. I did. There was a delay.

I had a question in terms of the

specific definition of the measure in terms of the numerator and the denominator. And thinking of the PSC as a clinical instrument but it's used as a performance measure is different from its use as a clinical measure. And in thinking about it in terms of there being a numerator and a denominator. And so I was a little bit confused in that under the -- in the numerator statement it says, in the sections that follow delineate specifications for two different meanings of each of these uses of the PSC. And then there's a list of the four different, I guess, numerator and then four different denominator statements.

And I was wondering if there was -- if you could say a little bit about the definition of the numerator and denominators, particularly in thinking about in the denominator statement the number of children ages three to eighteen receiving a well-child visit and then the number of children age three to eighteen seen for well-child visit in

a given measurement year. Is there intended to be some difference between those? I'm just thinking about the formal specifications that you're using for the numerator and denominator statement.

I think also this is sort of an issue that may come up in other discussions in terms of thinking about the use of a clinical instrument in -- as part of a quality measure versus the quality measure itself which uses the clinical instrument.

MR. MURPHY: So actually I'm not sure what the exact question is.

CO-CHAIR PINCUS: So can you get a little bit more specific about the numerator and the denominator of how, in terms of the actual performance measures that are reported?

MR. MURPHY: So start with the biggest and easiest one first. In any given year, the State of Massachusetts can figure out -- the numerator is all kids who had well-child visits, so I'm not sure which one of

these is which. But all well-child visits. So the process measure is whether the percentage of those kids that got screened and the number in the State of Massachusetts now is about 70 percent.

But other questions -- I think this is more where your question is going -- about the clinical outcome, you can also say, well, of the positive screens how many were referred, how many received services. And so you can track that, too. So one -- and finally, you can even look on a granular level or clinical level to see whether the kids that were positive last year are doing better this year.

I still don't know if I'm getting what the question is.

MEMBER SUSMAN: So from an NQF -- and maybe this is a staff point of view -- in general we've had this, there's been a linkage of -- for let's say a depression measure like the PHQ of doing the measure and then perhaps

following it to remission or demonstrating that there has been some follow-up, did the measure get repeated again. This, as I understand it, is proposing to do all the things together within this measure which leads, I think, to some substantial confusion when we start commenting on it because we're trying to consider a sort of prevalence or incidence measure around how many people in a denominator population get screened, and then how many get followed up. And is the follow-up actually improved or worse?

And by confounding all three of those issues together, methodologically and then just from a pragmatic viewpoint of our doing an assessment, it gets rather complicated.

MR. MURPHY: You know, I think probably other committee members can speak to that better than I can. I mean, it's certainly something we wrestled with a lot when we tried to put the PSC information into

the NQF format. What we eventually did is to say --- we can't say is it a process measure and outcome measure. We can only report on the ways in which it's being used and that people are using it in both ways.

And I noticed in preparing for this meeting that the very last measure you're going to discuss on the second day has to do with the multi-dimensional mental health screening assessment, and those measure developers are advocating for routine screening of adult medical patients for a broad range of psychiatric problems. So you know, one use of this is certainly a process measure for whether mental health problems are screened for in primary care. But again, the measure is also used in other ways, too. So how that fits into the NQF approach I'm not sure.

CO-CHAIR PINCUS: I think in terms of evaluating the measure on the criteria we should, you know, wait for Tami and Mike to

present on that. But one thing, though, that would be clear and, I think, helpful for the measure developer to give us is a greater degree of precision in the specification of the denominator and the numerator as it's applied and reported.

But are there any other questions for the developer?

MEMBER ATKINS: Well, I guess it's a follow-up question. But couldn't it be divided up into multiple metrics? You're asking all the right -- from a health plan perspective you're asking all the right questions, because people do screenings and then don't do anything with it. That's common. So instead of doing it for to get -- so they don't get in trouble, not because it's going to make a difference.

So once they do a screening, I want to know did they actually look at it and do something with it, and then did they do something useful? And all three of those are

critically important questions when you're looking at the actual benefits to the human being. But it may be that this is three different metrics and not one metric because you're -- and maybe that's a solution to the confusion.

CO-CHAIR PINCUS: And just as a maybe a recommendation to NQF, there should probably be a kind of a template that's developed for the use -- for how to describe and define quality measures that rely on clinical instruments as outcomes so that there's a standardized way of doing this. But anyway, we should probably move to Tami to talk a little bit about her evaluation, particularly of the criterion number 1, correct?

MS. FRANKLIN: Correct. Thank you, Harold.

And I just want to note that in our person and family centered care project, we are looking a lot at instruments and have

really done a lot of work around providing some guidance about how these measures should be constructed in that forum. So with that, Tami, if you wanted to walk us through importance?

MEMBER MARK: Yes. I mean, in terms of importance, there's clearly a large prevalence of undiagnosed, untreated pediatric behavior health problems that this measure tries to address. And you know, the measure -- as was mentioned by Mike, the measure is widely used in Massachusetts. Part of that stemmed from a lawsuit that, you know, identified a lack of screening and treatment and identification. So I think our -- you know, our internal committee felt that it did meet the requirements for importance and it performs -- our internal committee did feel that this measure met the criteria for measuring -- addressing a performance gap.

MS. DORIAN: Anybody else? Who is the second?

CO-CHAIR PINCUS: Mike?

MS. DORIAN: Mike?

MEMBER TRANGLE: Yes. I think what -- when we talked about in our phone call for the pre-group discussion, I don't think anybody thought it was unimportant. You know, I think the confusion about what it's measuring and utility of it wasn't even expressed as strongly as it was here.

In a lot of ways I think people are impressed with how it seems to have kind of started at a grassroots level. I didn't know it was because of a lawsuit. But how rapidly it --

MEMBER MARK: That's grassroots.

MEMBER TRANGLE: That's grassroots. You know, it could be on fire, those roots. But how rapidly it spread and how broadly it seems to spread. And at least from the comments we heard it looked like there was clinical utility for actual pediatricians and/or, I don't know, child psychiatrists kind

of in the exam rooms.

You know, not being a pediatrician I wondered but didn't articulate that I know the follow-ups are supposed to be at well-baby visits and those regularly scheduled visits. I have no clue whether it actually happens then. And if you're looking at evaluating it as improvement with all different kinds of time intervals in there, how that plays out as well, in terms of reliability. But we didn't really talk about that.

Are we going through one set at a time? We just finished importance and I rambled.

MS. FRANKLIN: Yes. No -- yes, and delineating the evidence, performance gap and priority, which I think you covered, both of you.

CO-CHAIR PINCUS: Any comments by other members of the committee?

(No response.)

CO-CHAIR PINCUS: Questions?

MEMBER MELNYK: I think it goes back to pediatric primary care. If we screen and we find, do we have the services that are going to be able to deal with it? Because on the United States Task Force, when depression screening came about, we changed that recommendation. We put a proviso in there, screen 12 to 18 year olds when systems are in place to accurately diagnose and treat. And I think it gets back to your earlier comment that you were making about screening and then what do we do about it, if it's positive?

CO-CHAIR PINCUS: And just let people put their things on the side so that we can see. So I think Rhonda was next, then David, then Larry.

MEMBER ROBINSON BEALE: This is really a clarification question. The pediatric screening, the PSC, when it's administered, is it something that's based on clinician judgment or is it based on patient input?

MEMBER MARK: Yes, I think it's a parent, it's given to the parents.

MEMBER ROBINSON BEALE: So my question would be what testing has been done? I just don't know the tool to look at the reliability of the tool, particularly when we're talking about measuring improvement.

CO-CHAIR PINCUS: Rhonda, we're going to get to that criterion later. The way this is -- you heard at the beginning where we said that we're going through criterion by criterion, not overall. So we'll get to that.

David?

MEMBER PATING: So my question goes to the issue of the relevance. I wasn't quite sure whether the need is better to find as we need to do screening or that we need to use the PSC. I think it's similar to maybe what you were measuring. I believe that there's competing measures in the field. I don't know their revelative breadth of them or the use of

them and so if we're driving with one measure as the goal again, or the gap is that nobody's using the PSC and that's standardized, or nobody's screening, which is to me just slightly different.

So I wasn't quite sure in reading this the gap. I understand there's a screening gap but is this PSC the only way to fill it and is there another way to define the measure that would address the need more broadly?

(Inaudible comments.)

CO-CHAIR PINCUS: So this is an issue of how we approached this criterion. So just to be literal about it, it's evidence to support the measure of focus. So this is the, you know, extent to which the measure is focused and sort of a concept is appropriate, not specifically whether or not it's, this particular instrument is the most valid of all.

And so as we go through this, you'll

see that when we -- there's kind of a method to the madness here, in terms of how one looks at each of the individual criteria. Because later on we'll do some work around looking at the comparability of this measure as compared to other measures, whether it's sort of best in class or not.

So the focus for this criterion is really focused around does this concept of measurement, is it important, okay? And we'll get to sort of the more comparative stuff and also the reliability and validity of it sort of within other criterion. So that's what we're focusing on right now. So you know, basically with regard to this measure, you know, is it important to screen? And it looks like this is actually a composite of several different measures, and something that we may want to get back to in terms of further clarification. Is it important to screen kids, number one, and is it important to follow up and see whether they improve?

Because it seems to me that's part of it. And so that's the concept. Whether this is the best or not, we'll get to that a little bit later. Okay?

Other questions? Comments? So David and then Jeffery.

MEMBER EINZIG: So just wanted to give the clinician perspective on doing screening if there aren't necessarily resources to refer to. So in our clinic in Minnesota, you know, that was an initial concern there, that you're going to be screening all these kids and what are you going to do with them now that you've screened? But you know, I think it's turned out that, you know, these kids do ultimately get referred if they deserve referrals.

And comparing it to, say you don't screen, one of the problems I have as a child psychiatrist is I get kids who are worst of the worst because they weren't referred to me sooner. And so if you're not screening and

the kids get worse down the road, I think the purpose -- it is important to measure with the screening tools so at least it starts the conversation and some of the information gathering for the clinician and families.

MEMBER SUSMAN: This may be a small quibble but a system that was doing very poorly at getting its kids into well-child exams, all those children are excluded as I read it from the denominator, and may be a very poor system compared to one that is making outreach efforts to get all the most challenging children into the population for well child checks and therefore screened.

If the measurement period is long enough, a year, one would assume at least in early childhood that might be a minor issue. But as you get more toward the adolescent age, I think it could be a important problem in considering one plan's performance versus the other. So perhaps it would be helpful if the measure developer has any information on

differential attendance where you know what the true denominator is and then look at the actual screen denominator.

CO-CHAIR PINCUS: Just a comment. So again, it's a little bit ambiguous. One of the problems with the criterion is that, as we're going through this now, you can see is that everything adds up to importance, you know? So you can talk about almost anything with regard to importance along the way, if it's not feasible, if it's not important --- so that's one of the issues to think about. And as we go through this we need to think about how we sort of split out some of these.

But Jeff, I think you're right, and so one question is, at some point we may want to come up and bring a suggestion out about the possibility of having a balancing measure for this that looks at the proportion of people, of enrolled children within these age groups who have had a well child visit. And I don't know whether such a balancing measure

actually exists or not but it's something to think about and talk about with the measure developer.

Bob?

MEMBER ATKINS: With regard to the issue of once you screen somebody does your delivery system or does the -- the local area have the ability to help that person? Again, I'm looking at things through the filter of a health plan. That's just as important to me because if we're not driving -- I mean, that's a huge problem almost everywhere.

So we need to actually know where that exists as part of -- if you will, as part of the motivation for change. Because if we don't know where there are gaps in service there's nothing we can do about it. And it also motivates delivery systems to say, well, we just don't care about that. Well, now that we're measuring it, maybe you ought to. So I think actually the lack of services increases the importance of this metric.

CO-CHAIR PINCUS: Rhonda?

MEMBER ROBINSON BEALE: I want to echo what Bob said, coming from a health plan perspective. I think given this time and the changes that are occurring in healthcare where the redistribution of funds and interests is on the radar screen, I think it's very important to be able to quantify the areas where there is a need for services and the existence of issues. And I think we all would agree that with children, we've kind of missed, well not kind of -- we have missed the boat.

And so having a measure that can -- and I would really specify something that Jeffery was talking about, that it should be a measure that looks at the entire population of children, whatever that entity is responsible for, and that measure be used for the entire population so you can understand the prevalence of the issue. I'm not saying about what it's describing but the prevalence

of the issue which I think is exceptionally important right now.

CO-CHAIR PINCUS: Well, I think we're ready to vote on this criterion. And let me try to summarize.

It sounds like that there is some degree of consensus that this is the concept of measurement, screening by using a clinical measure and also follow-up and assessment of improvement which seems to be part of the composite of this measure is an important issue. That there is -- this is a highly important issue in terms of impact on kids and families, that it's -- there's a gap because there's a lot of people don't screen and don't assess and don't follow up. And that you know, that there's also a gap in terms of measures for kids. But that there's also a number of issues that have come up that we'll deal with also, that will come up along with some of the other criteria. And there may be some suggestions for balancing measures or for

other ways of improving the measure.

So we're ready to vote?

MS. BAL: So we're going to do a test run, hopefully everybody's ready to work. Please aim towards me, not the screens. And so just a little background. So the voting will work if we have above 60 percent, that means that's the decision you made, the consensus reached. If it's between 40 or 60 that means consensus was not reached and we'll put it in the gray zone. And then if it's under 40 percent that means consensus was not -- I'm sorry, that means that the measure failed that criteria.

So for the first one for evidence, so for the numbers next to whatever the decision point is, so if you think the measure rates a high you put a 1, moderate 2, low 3 and so on. And then that's how you vote. You only need to push it once, you'll see a green light indicating that you have voted and it was registered. And if you see a flashing

green light, that means the batteries may be low. Please let me know if you do see a flashing green light. And the vote, it won't work right now because I haven't started the vote. So don't fear, for right now if it's not working. Also if you get a red light that means it didn't register and definitely let us know. So we're going to do a test run real quick.

CO-CHAIR PINCUS: So just to clarify, we're voting on each of the sub-criterion?

MS. BAL: Yes.

CO-CHAIR PINCUS: Okay. So it's not just one gemish for the whole criterion 1, we're looking at each 1(a), 1(b), 1(c)?

MS. BAL: Yes. So this is only evidence -- well, right now let's just do a test run, don't -- you can put your real vote if you like but just as a test run to make sure everything works. Maybe you could actually -- everybody could just hit 1 to make

sure that works.

Give me one second. Once I hit the button then you'll be able to.

MS. DORIAN: So you're looking at the two side screens up there?

MS. BAL: Yes. But make sure you're clicking towards me. Yes.

And so when you see this clock you'll have one minute to vote. And if you see -- please let me know if you have a flashing green light or a red light. Just the number.

MS. DORIAN: And if you change your mind you can push another number. It will only register the last number that you actually --

MS. BAL: The last one. You can change your mind.

CO-CHAIR PINCUS: So the clock is ticking?

MS. BAL: Yes. We want a green light, no flash or red.

So we only have 22. Could everybody just try to aim at me again and just make sure you get that green light?

CO-CHAIR PINCUS: What is the denominator?

MS. BAL: We're trying to find 24.

(Laughter.)

MS. BAL: Are we at 23? Oh, time's up. Did everybody get a green light when you selected it? Hold on, we're just going to double check that we have 24.

Okay. Maybe one more test. Sorry everyone.

So we're going to try one more test and, just push it once and then we're good to go, hopefully. Oh, sorry, wrong button. Okay, ready now, go ahead. This is a test run, not the actual vote.

MS. BAL: We're at 23 -- 24. Okay, we're good to go. All right, everything's working.

CO-CHAIR PINCUS: Okay. So now

we're going to do it for real.

MS. BAL: Yes, now we're going to do it for real. This is the real vote.

CO-CHAIR PINCUS: And again, we're voting on 1(a), the first of three different criteria.

MS. BAL: Yes. Okay, so go ahead and vote now. We're missing one person, if we could just try one more time. Just click whatever your answer was. Oh, that makes sense. Thank you.

Okay. We have for evidence for 722, we have high 16, moderate 7. Zero low, zero insufficient evidence and zero insufficient evidence with exception. So we will move forward to the next vote.

CO-CHAIR PINCUS: Would you co-chair with me on this committee?

CO-CHAIR BRISS: Good morning. Thanks to everybody for being patient with my Atlanta to D.C. commute this morning. And I'm Peter Briss and Ann's not here anymore but I

have no conflicts.

MEMBER GOLDSTEIN GRUMET: Good morning, I'm Julie Goldstein Grumet with the Suicide Prevention Resource Center. I also live locally, I had to get my kids to school and then drive through rush hour so I apologize.

MS. DORIAN: And Mady?

MEMBER CHALK: I'm Mady Chalk from the Treatment Research Institute. Sorry to be late, no conflicts.

CO-CHAIR PINCUS: Okay. So now we're going to be voting on 1(b), the concept of a performance gap.

MS. BAL: Okay. Voting for gap is now open.

Just one more time -- sorry. We're at 22, we just need one more vote.

Yes -- oh, we got it. Thank you.

So for 722 gap, the final results are high 19, moderate 3, low 1, insufficient zero, and we will move forward with this

measure.

Are we ready to vote for this one, too?

Okay, sure. So we are now voting for high priority, which is -- addresses the specific national health goal priority or data demonstrated a high impact aspect of healthcare. The options are 1 high, 2 moderate, 3 low and 4 insufficient. And voting is now open.

All right. Thank you everyone.

And so the result for impact for 722 is high 20, moderate 3, low zero, insufficient zero, and we'll move forward to the next criterion.

CO-CHAIR PINCUS: So now, Tami can you introduce the discussion about scientific acceptability and going through each of the criteria?

MEMBER MARK: Yes. So first we're going to discuss reliability. There's a lot of information that was submitted on this. A

summary is provided on page 41 and there's also a large summary document of all the literature related to the pediatric symptom checklist that was included in the submission.

So essentially, the reliability of the PSC survey instrument has been repeatedly demonstrated, and you can see that in the references that are provided. However, the reliability of the PSC as used in administrative claims, which is essentially what this is proposing to do, to use it at a system level, has not been formally assessed as of this writing, but reading from this, indications of the reliability in the Massachusetts Medicaid claims data can be inferred from the stability of the rates of positive screening from quarter to quarter in the statewide CBHI data over six years. So I'll leave it at that.

CO-CHAIR PINCUS: Could you explain that again, and explain what you interpret that to mean? I'm a bit lost with that.

MEMBER MARK: Yes. So I think what they're saying is, as an instrument, if you just take the PSC and you do your standard Cronbach Alpha, it shows high reliability, but it's a little trickier when you think about how you do it in a -- how do you test the reliability in a office-based setting where it's based on claims data? And what they're arguing, I think, is that if you saw a lot of variation in the prevalence rates over time, that might indicate a lack of reliability. But the fact that the overall prevalence rate seemed to be pretty stable suggests that it's reliable. And I don't know if the developer is still on the phone and wants to speak to that, but that's my interpretation.

MS. DORIAN: Michael?

MR. MURPHY: Yes, I am still on the phone and I'd be happy to speak.

MS. FRANKLIN: Go ahead.

MR. MURPHY: And I thought that was a great summary. As we move up from the

individual case to the system level we don't have a lot of the kind of data we wish we had. And the example you gave is one way we infer it. We do give it repeatedly at Mass General Hospital every three months, it's not the same as doing it in pediatrics. We find good reliability there. And we actually have done some case reviews in clinics where we have access to the EMR and good reliability there. But certainly not much published yet.

CO-CHAIR PINCUS: And in validity?

MR. MURPHY: Yes, again --

CO-CHAIR PINCUS: I'm asking Tami.

MR. MURPHY: -- I think we've got validity on both the process measure and the outcome measure. Some of the work we've done in Chile where they use it as --

CO-CHAIR PINCUS: Actually I was asking Tami to continue.

MR. MURPHY: Oh, I'm sorry. I'm sorry.

MEMBER MARK: I will parse this by

saying I think the point that was brought up earlier about the fact that this is multiple measures rolled into one makes the validity discussion a little challenging because there's validity and then on top of that you have this validity of the instrument again, the PSC, which there is a fair amount of data to show that it's valid. You know, it measures what it purports to measure. But in the system level, you know, a little more complicated.

So the -- I think that perhaps the trickiest point about validity is the validity of the outcome measure. And you did -- the outcome measure again is, did someone who was screened, identified on the PSC at one visit show a reduction on the next visit and is that a valid measure of improvement of quality of care?

And in the submission they highlight that the American Academy of Child and Adolescent Psychiatry does, you know,

recommend this kind of routine screening as a way to improve outcomes but that the U.S. Preventive Services Task Force did not find sufficient evidence to recommend routine global psychosocial screening of school age children or teens at the present time. And we had a discussion as well about the evidence from the Hacker research as to whether the instrument as implemented in Massachusetts found an improvement in outcomes. And I don't know, I think the developer was going to provide some additional information on that.

The publications to date do show that the screen has led to improvements in case identification and into referral but there was no evidence to date that it led to improvement in outcomes.

CO-CHAIR PINCUS: We'll get into the developer's response in a minute. But first I want to ask Mike to just comment as well.

MEMBER TRANGLE: That was such a good summary I don't think I have anything to

add.

CO-CHAIR PINCUS: So let's ask the developer if he could sort of respond with the information that was provided to NQF subsequent to the workgroup call?

MR. MURPHY: Yes, thank you again.

And this came up on the workgroup call, you know, the lack of U.S. Preventive Service Task Force endorsement, it's certainly true. But there are certain things that establish their validity because of expert recommendations like AAP or the American Academy of Child and Adolescent Psychiatry, or the fact that EPSDT, the national program, recommends it. So -- and I think Tami also mentioned the multiple domains of this measure process outcome make it necessary, ultimately, to do a validity assessment of all of those things.

But as Tami also said, the Hacker stuff does show that routine screening on a large scale leads to increased identification

of kids with the problems, increased services and some of our work in Chile shows that kids who get services, do better on the PSC and kids who do better on the PSC have better grades and better attendance and large longitudinal datasets. So this is a measure that's on the bubble, it's not going to be in everybody's endorsement package, but we're hoping that NQF sees it as something important that can be recommended and endorsed at least by NQF.

CO-CHAIR PINCUS: So why don't we open up for discussion by the committee. And just one question for -- in terms of the NQF, I have a clarification. You say a little bit about any changes in the scientific acceptability criterion in terms of how one applies the issue of endorsement by an expert group versus, you know, what sort of empirical evidence is necessary.

MS. FRANKLIN: For validity? Are we

looking --

CO-CHAIR PINCUS: For validity and reliability, yes, especially validity.

MS. FRANKLIN: There have been --

CO-CHAIR PINCUS: That's -- because that -- I know there's been a lot of discussion about, you know, sort of face validity and expert opinion recommendations versus empirical evidence from studies.

MS. FRANKLIN: So there hasn't been a change in our criterion, per se. We do have more stringent guideline -- guidelines about how to look at face validity. And at each of your places you'll have an algorithm that kind of walks you through how you should look at the validity criterion. And you -- yes, Harold has it in his hands.

And this is the guidance that we're going to be using as we look at the validity aspects of each of the measures. I hope -- does that help answer your question, Harold? Or is there --

CO-CHAIR PINCUS: So yes, maybe you just want to --

MS. FRANKLIN: You want to walk through?

CO-CHAIR PINCUS: Yes, walk us through. Be clear because this applies to all of the measures so it would be useful. And this is, I think, going to be among the most critical criteria that we discuss, so it's good to be really clear about this.

MS. FRANKLIN: So specifically looking at the guidance for -- and I'm looking at the guidance, algorithm number three guidance for evaluating validity. Is that your specific concern?

CO-CHAIR PINCUS: Mm-hmm.

MS. FRANKLIN: And first you're looking at whether the measure specifications are consistent with the evidence, and I think we just had that discussion about linkage back to evidence presented by the developer in Section 1(a) of the measure submission form.

And you want to look, and make that evaluation as a committee member, as to whether the specifications are consistent with the evidence presented.

And we did hear from the developer that they submitted some additional new information in the evidence realm. If you find that there isn't a clear linkage back to that evidence to support the measure you would rate this measure as low. If you did find that it was consistent, evidence consistent, you move on to whether a question of whether the potential threats to validity have been addressed by the developer, and you'd be looking at the exclusions provided in the measure. You'd also be looking to see if the measure really was able to identify differences in performance among the measured entities, and you'd also look at any ways that missing data was handled to reduce bias.

If you find any of these issues not addressed or of concern, you would rate the

measure as insufficient. If you were satisfied with this -- had a satisfactory answer to that question, you move on to look at whether empirical validity testing was conducted in this case, or you look at whether face validity was presented to support the measure. And in the case of face validity, you'd only be able to rate the measure as a moderate.

And I think that's, you know, where we -- this is a critical pieces of this particular algorithm for this measure. Are there questions about how you would walk through the validity piece for this measure?

CO-CHAIR PINCUS: No.

MS. FRANKLIN: Hearing none, Harold?

CO-CHAIR PINCUS: So let me sort of call this in terms of -- people who want to speak --- put their things up. So Bob, Mike, Jeff and Larry, and Bernadette and Peter.

MEMBER ATKINS: A couple points. I guess, given the state of behavioral health

services, I'd be pleasantly surprised if a screen measure was tightly linked to ultimate clinical outcomes. So I don't see that as a particular challenge to the validity of a screening measure. I think that that's the ultimate goal but that's an end state. So that would be one point.

I do have a concern about the validity. Second point, I have a concern about validity with regard to referral to a behavioral health clinician. Increasingly -- I mean, we know for a long time most of the behavioral health services in the United States are delivered in the primary care setting not by behavioral health clinicians. And I understand that there's been a longstanding concern about whether or not the quality is what we would wish, frankly in either setting, but that's a question.

But I do think that the other thing that, where I question validity of using referral to behavioral health clinician is

with the increasing sort of growth of integrated practices, whether it be a patient that's in a medical home or a behavioral health home where there are behavioral health clinicians within the practice. And that might not look like a referral, it would be contained within the primary care setting. So I'm not sure that that -- how to address that within this metric.

I'll stop there.

MEMBER TRANGLE: It's interesting, my thoughts were similar but not identical to yours.

I think we should think about, you know, this is a multi-functional potential measure. You know, at least in my mind, it was clearer if I tried to separate out the functions and think about how valid it is, you know? So in my mind's eye, I think about how valid it is for picking up and screening, it strikes me as reasonably valid.

When I thought about how valid is it

for referrals, you know, they're your points about is it always going to be a referral and how is that defined and how is that captured? And even if you were capturing it accurately, the other question I have is, when you're making a referral to a behavioral health person, about 50 percent never show. And if all you're measuring is the referral, you're missing a lot of the action, you know?

And then the part that I would probably have the most questions about is really sort of improvement, you know? Is it valid for improvement in terms of remeasuring and seeing if the numbers go down, if it was voodoo that did it or five other things? You know, who knows what the factors might have been, and it wasn't the referral.

So I think we should almost -- we should think about the possibility of separating this measure into different, three different parts and voting separately. I don't know if that's kosher, you know, in

terms of the methodology here.

CO-CHAIR PINCUS: Jeff?

MEMBER SUSMAN: I think you've articulated much of the concerns I had at the beginning of this discussion which is essentially we have, one, a measure to find prevalence or incidence, depending on how you're going to use this. Two, to find some change for process improvement.

The question on the process improvement side I had is do we have enough data, and this might go back either to those of you who reviewed more in-depth, or the developer, do we know what a substantial or significant clinical change in this measure is? And what data is the validity and reliability of that change based on? I was poring through the documents and I confess I didn't see strong evidence in that arena but maybe I missed it.

CO-CHAIR PINCUS: Bernadette?

MEMBER MELNYK: Because the U.S.

Preventive Services Task Force came up again, I just wanted to give people insight because I was on the task force. And actually brought forward this topic recommendation, because I feel it is so critical because we have one out of four children now with a mental health problem, yet less than 25 percent get any treatment. The task force voted this as a high priority for topic review, but it never made it to topic review because of other priorities. And because of people being concerned that, again, we screened and we find and we have interventions.

The Task Force had a lot of concern that we were taking on topics that again we don't have sufficient evidence for. And there was a lot of criticism about the task force continuing to put out insufficient evidence recommendations. So I just want to give you the background on that. There have been more recent studies that have shown screening with this particular instrument. There are

evidence-based interventions to improve outcomes. And those are more recent in the literature. So again, I just want to advocate for, this is a solid, valid, and reliable tool that more recent studies are coming out that say we can do something of value if we find it.

CO-CHAIR PINCUS: Rhonda?

MEMBER ROBINSON BEALE: I agree with the previous statements that I think it's difficult to really vote on this as a composite measurement. I think it would be doomed to fail in that regard because I think there is still hold.

I do agree that it's very important, it has significance, and I think that is, in itself, probably what drives the interest in this particular measure. I'm not clear that the PSC is anything beyond a screening tool. I don't see it as a assessment tool, and I tried to look through the studies and I certainly wasn't clear that it was a tool that

showed validated improvement. And so I really find it hard to go further in terms of really looking at this as a multiple measure making one vote as opposed to splitting it out in some way.

CO-CHAIR PINCUS: David then Larry then Peter then Mady.

MEMBER EINZIG: So I'm a simple guy, I just want to try and conceptualize this. So if this is a screening tool that's simply used to identify early issues with kids, I think that it's -- you know, I think it's great. I guess I need a little clarification as, is this -- if the intent of this measure is to be used as a quality measure in terms of outcome, in other words, is it good quality if the score goes down as opposed to up, you know, that's where I have a problem with it. But if it's just a simple quality measure of are we screening and are we doing something about it, that's where the greatest validity is with this tool.

MEMBER MILLER: In addition to simple, I'm also old and paranoid and suspicious and I have an issue about the reliability, I believe.

As far as I understand the administration is a 35-item parent administrative checklist, parent or youth. I don't see anywhere that the clinician also does this or validates it in some ways. And the reason I say that, we had in our system in Arkansas, we've used the Youth Outcome Questionnaire. And when we only had the parent fill it out, sometimes there's an incentive to make the child look better or worse or whatever. Some parents are concerned that if they make their child look worse or accurate that they're going to be taken away or something like this. And so when we added an element of having a clinician fill it out, the scores changed dramatically.

And so I'm concerned about just the basic reliability of having a parent fill out

the thing if a clinician hasn't also filled it out and sort of reviewed it, and especially in terms of the outcome improvement. So that's a concern that I have.

CO-CHAIR PINCUS: Peter.

CO-CHAIR BRISS: So on thing, all of my things are reacting now to people's comments. So it's -- one of them is that I would be a little careful throughout the whole two days about giving behavioral health interventions a relative pass. I think that the -- it's sort of the standard for the evidentiary standards for almost anything, as does screening and some associated through improved outcomes. And if it doesn't, it's probably -- it may not be -- and especially for a measure like this one that's likely to be used in a variety of settings, including primary care, right? Where people have lots of other pressures on their time.

And then I would consider this as four separate measures. I think it's

impossible to think about reliability and validity of this measure without thinking about it one thing at a time. It probably -- in my view, it probably could be reliable and valid if it's a measure of "did screening occur," as sort of a process measure of "did screening occur?" I'm not sure that the developer has made the case here that it's actually reliable and valid for any of the other three proposed uses. And especially for the "did outcomes improve," both because at least in the submission, as I read it, you know, unless there's new data that wasn't actually submitted, I saw very thin evidence of outcome improvement. And to the extent there might be outcome improvement, this kind of a measure is very subject to regression to the mean.

And we talked to the developer a little about that in the workgroup call, he said it's about 50 percent. But without better data and evidence presented, I think

it's very hard -- beyond what's presented in this submission, I think it's very hard to evaluate its potential.

CO-CHAIR PINCUS: Mady.

MEMBER CHALK: I raised the question that Peter raised about to what extent, because there are other measures that we're going to be talking about that involve screening that are identified as process measures primarily. And I was curious about why this was tagged as an outcome measure specifically when, as a process measure, in my view, it works just fine for right now, but not as an outcome measure.

CO-CHAIR PINCUS: Tami.

MEMBER MARK: I think part of the concern that I have, and maybe it's underlying this discussion, is that the screening in itself may not -- screening and increased identification may not lead to better outcomes. In fact, you know, it may lead to worse outcomes as people get inappropriate --

children get inappropriate medications and get iatrogenic problems. And we're talking about implementing this on very large populations so the potential for, you know, having worse outcomes across many, many children is a little concerning. But at the same time, we do recognize that there is a significant need in this population. And if this were valid and did result in improved outcomes, it has the potential to be very important.

So given that, is there a way to vote on it, you know, giving it a contingent vote saying, you know, we would like to see this but, you know, you need to come back with this data showing this actually is going to improve population level outcomes? Or do you just -- you know, it's a "yes" or "no" and then three years they come -- they can submit it again? You know, what are our options here?

MS. FRANKLIN: So we -- first of all, it may not be exactly three years. But

we have to vote on the measure as it's constructed currently before us and we would not be able to -- we could make recommendations for the future on the measure. But we'd have to look at the measure as it is and evaluate it on its current measures -- merits.

CO-CHAIR PINCUS: So what if it was not accepted but we made recommendations, when would they be eligible to come back?

MS. FRANKLIN: As -- we don't have a specific timeline right now. But it typically could be up to three years but it could be sooner, depending on if there is new evidence that they've presented to us.

And one other note, if you were to -- and I'm not sure if this is an option given the way the conversation's going, is if the committee decided to go forward with this measure and recommend it and evidence changed, the developer could come back within the year with an annual update with any material

changes to the measure for an ad hoc review. So that's also an option. But that's for the committee's determination.

CO-CHAIR PINCUS: So I want to make an observation as chair and then sort of a comment as a committee member. So as I see it, one of the issues that the -- our discussion suffers from is that there really isn't a clear specification of the performance measure for the use of this clinical instrument. And that for other clinical instruments that have been proposed, they actually come in a form of several different measures, each with a specific sort of assessment of the evidence and characteristics and all the other criteria for those specific uses. And I think that that's creating a problem in terms of how we discuss this. And again, I would urge NQF to consider developing a template for those kinds of situations so that we can sort of look at it that way.

And just a comment from my own point

of view, so looking at this measure as presented, I really think it's problematic in terms of the scientific acceptability. In terms of, number one, the reliability, you know, as Larry mentioned, I'd be concerned about -- you know, and particularly Tami, in terms of this widespread applicability in terms of both false positives, false negatives and, you know, especially with the different -- the heterogeneity of how it might be used in clinical settings.

The fact that there is really, for at least some of the purported uses, there really is no evidence of linkage to outcomes.

And number three is, there also does not seem to be a formula for risk adjustment. And it certainly is likely that different settings are likely to have very different groups of individuals that are coming in with different risks and not -- it's not clear how one would adjust for that if this is being used as an accountability measure. So you

know, I have concerns and would recommend that the developer come in with a much clearer set of performance measures using this clinical tool with increased data.

CO-CHAIR BRISS: I just have a couple process questions for the staff.

So I'm -- a lot of people around the table have expressed varying levels of concern based on which of the uses we're talking about, right? And so do we have options of splitting out uses as a committee or do we have to vote on -- do we still have to vote on the whole set together and then it's -- you know, we -- NQF's been talking about fit for purpose, right? And so someday we're going to get to the place where we can say this measure can be used for these purposes and not those, but I don't think we're there yet, is that right?

MS. FRANKLIN: That is correct.

CO-CHAIR BRISS: And so the vote that we would have to take is for all of the

proposed uses, is that right?

MS. FRANKLIN: That's correct.

CO-CHAIR PINCUS: Jeff, do you have a comment?

MEMBER SUSMAN: Yeah, just briefly.

I share all the concerns, but this pediatric symptom checklist probably has a larger dataset than most of the things that we consider. But because we've got these three or four different uses sort of in a fruit cocktail makes it very difficult.

I wonder if the staff could somehow, with the measure developer, fast-track this back and -- to separate out those uses? Because I think our job would then -- would be a lot more straightforward and we could make some discernment about what uses we would be happy with the degree of reliability, validity and other uses that we might not.

CO-CHAIR BRISS: I have a question about that. This is just a process question.

MS. FRANKLIN: Yes.

CO-CHAIR BRISS: So if we were to send this one back for further work, it could conceivably come back for an ad hoc review in a year, couldn't it?

MEMBER KELLEHER: And to add to that, if it -- could we also make recommendations about whether it should come back as separate but paired measures rather than the composite that it is right now?

MS. FRANKLIN: You could. And I just want to be clear, we'd have to -- for it to come back for an ad hoc it would have to have been recommended by this committee for endorsement for it to be able to come back as an ad hoc. So that would be the catch.

MS. DORIAN: I did just want to add a little bit of context just to make sure we're consistent across committees. I'm also working on a person and family centered care project which recently evaluated PROs, or person reported outcomes, and so this is similar in a lot of ways. And this is

something that NQF has struggled with in terms of these kinds of measures. Do we allow multiple measures to come in on one form or not? You know, we have to consider the burden to the developer and also the fact that, as you all have said, that they're very different uses and so they do, theoretically, seem different.

And we did actually allow at the end of the day the developer to submit all of the measures in one form and the committee evaluated it as one measure because the sort of argument from the developer's point of view and the user's point of view was that these are always reported together. I'm not sure if that's -- I think that's the case with this measure. But I just wanted to add that little bit of context, that that committee didn't hold it against the developer necessarily, that it was -- it all came in as one rather than separate measures.

CO-CHAIR PINCUS: I guess just to

clarify, I wasn't necessarily recommending that it has to come in as separate measures, but certainly the different uses have to have much more clear specifications for numerators and denominators as well as the risk adjustment, you know, that's appropriate for the fit. It's got to show fit to all the purposes, even if it comes in as one measure.

MEMBER CHALK: I think this is too important a measure, given the lack of such measures, to just toss it. No, I'm not talking about voting for it, but toss it without making a statement by this committee, we've done it in other committees, that says what we think the importance is and why -- what we want the developer to do and how we want it to come back so that NQF does not forget that this measure is hanging out there.

MEMBER KNUDSEN: I have a question. So I'm a little confused in terms of our voting that's coming up.

(Laughter.)

MEMBER KNUDSEN: Are we voting on four measures, two process, two outcome, or are we -- this is not a composite measure because that's not what was brought forth initially. So -- or is this an overall measure that we're voting on? But then how do you vote on reliability and validity of four different things with one vote?

CO-CHAIR PINCUS: Well, I --

CO-CHAIR BRISS: Let me try that. Let me try to answer.

So I think that we have -- I think that the vote is we would have to be voting that this is reliable and valid for any of the four uses for which it's been proposed, right? And for all of -- in principle, for all of the uses for which it's been proposed.

MS. DORIAN: And as Angela had mentioned before, as you are a standing committee now, you would be the ones reviewing, making recommendations now, of course, but then reviewing any changes

subsequent to this meeting. The measure could come back for an ad hoc review if it was significantly changed.

MEMBER ATKINS: So could you explain the impact of this vote on the ad hoc timeliness of the term? Because I'm a little -- it seems like it might be unintended consequences that we're going to shoot this thing and we don't want to shoot it, we want it to come back.

MS. FRANKLIN: So what I heard was a recommendation that just looking at the four numerator statements, as it were, the committee wasn't feeling comfortable about the reliability and validity of all of these being used. And that's what the decision's about, on the reliability and validity vote now.

And with -- the committee would have to recommend the measure going forward, or the other option, with certain specifications that would need to be changed by the developer. But at this time we probably would not be able

to have the developer make those changes in a timely way. So -- in this project.

You'd be also -- you'd also have the option to not vote for the -- not recommend the measure for endorsement and the developer could bring back this measure differently formulated as instructed by the committee for a full endorsement review again. And that's really kind of the two options that we have.

MEMBER CHALK: So I have a question. Could it be recommended that the developer view this as a trial measure?

MS. FRANKLIN: That is reserved actually for electronic measures.

MEMBER CHALK: E-measure specifications?

MS. FRANKLIN: Yes.

CO-CHAIR PINCUS: I guess let's hear from the measure developer before we vote?

MR. MURPHY: I'm trying to stop crying and I'll get myself together just for

a second.

Actually, I thought this was a great discussion, and I think some of the ways -- I'm also quite heartened that some of the ways out of the -- I totally agree that the validity and reliability of this for some of the other uses just hasn't been demonstrated, whereas the validity and reliability as a process measure is probably pretty solid. So there may be some need to break it out and we're certainly willing to work with your committee in any way.

The only thing that -- a couple things concern me, and I think it just reiterates -- somebody just said we don't want to kill it, we want it to come back in a different form. I don't know if there are any pediatricians on the committee here, but I think this is one of the few pediatric, the only pediatric mental health measure. And its virtue is that it brings mental health into pediatric primary care with all the lack of

specificity so far. So I would hope that we find a way to keep it alive but buff it up a little bit so that it can be looked at and voted on appropriately.

And by the way, it is about to be e-specified, if that helps. SAMHSA just awarded a contract to have it e-specified by Mathematica, so they're expecting e-specification pretty soon.

CO-CHAIR PINCUS: So yes, we're now -- I think that we need to vote. I think that -- you know, I think that everybody here feels the need for some kind of way to address this issue in this population. And I think, on the other hand, you know, there -- we do need to adhere to the -- you know, to the process and apply the criteria in a -- you know, in a uniform way across all of the measures in terms of fairness and also in terms of -- you know, and I'm also sort of going back to what Tami mentioned about the potential risks in terms of application and its use in ways that

might actually create harm. And so that we need to think about that.

So you know, I think we need to vote on the measure as proposed, applying the criteria as expressed by NQF. I think we can also accompany that recommendation -- whatever recommendation comes out of the voting, accompany that with a very strong statement about the clinical need for this, its importance in terms of the needs of the population and addressing that, and some specific recommendations for how this measure, or any other similar measure, should come back to us in a way that would be more acceptable and that also could be done quickly. And that, you know, even though NQF has certain procedures, I think there's certain ways in which they can work to kind of bring it back more quickly if we make that case.

CO-CHAIR BRISS: The only other thing that I'd add is that the point of standing committees is to make things faster

in review and so we don't have to do what we would have used to have to do, which is reseat a new committee and come back in three years.

The other thing that I would say is that part of the point of the process is to improve measures, right? And so we'd rather have a better measure in a year than have a measure that might be misinterpreted or even harmful soon.

CO-CHAIR PINCUS: So let's -- Poonam, can we proceed to voting?

MS. BAL: Okay. So we'll be voting for the reliability for 722 which includes the specifications and testing. Voting is now open.

Oh, I'm sorry. One high, two moderate, three low, four insufficient.

Okay. So the results for reliability for 722 is high one, moderate four, low three, insufficient fifteen. And we don't move forward with this measure after this vote.

CO-CHAIR PINCUS: Okay. And so will the recommendation be accompanied by, I think, a statement very similar to what Mady spoke of earlier?

MS. FRANKLIN: Yes.

CO-CHAIR PINCUS: Any other final comments?

(No response.)

CO-CHAIR PINCUS: Okay. Let's return to the measure -- yeah, our agenda and measure 0108. And Peter will lead us in that.

#0108: ADHD: Follow-Up Care for Children Prescribed ADHD Medication (NCQA)

CO-CHAIR BRISS: So this is 0108 follow-up for children provided ADHD medication and David is the lead discussant. So if you could kick us off?

MEMBER EINZIG: Okay. Measure 0108, follow-up care for children prescribed ADHD medication, so kids not -- with ADHD can --

CO-CHAIR BRISS: I'm sorry, I misspoke. Can you let NCQA introduce the

measure for us?

MS. HUDSON SCHOLLE: Good morning everyone. I'm Sarah Hudson Scholle, I'm Vice President for Research and Analysis at NCQA and I'm here with Junqing Liu who is our research scientist. And we're delighted to have a number of measures for your review today.

The first measure is this measure that looks at follow-up for children who are on an ADHD medication. So this measure is based on claims data. It's been around in our HEDIS measure set for health plans since 2005. It's currently used in a number of federal and state programs including the Children's Medicaid Core Set, PQRS, Meaningful Use. It's proposed for later stages of reporting by the Quality Rating System for Exchange Plans.

It is a claims-based measure so the purpose of the measure is to say if children are on a medication for ADHD, they should be getting appropriate follow-up. So it's

looking specifically at that construct. And while we have other work underway to try to look at outcomes for children with ADHD and other considerations, this is one that's currently in use because it's feasible from claims data. So we're looking to see whether children have follow-up visits to monitor their response to treatment and any potential side effects and with the goal that this medication management will support better use and outcomes.

Thank you.

CO-CHAIR BRISS: Thank you.

And now to David.

MEMBER EINZIG: Okay. So follow-up care for children prescribed ADHD medication. I think this is worth mentioning that this isn't measuring kids with ADHD on medication but it's just simply kids who are getting ADHD medication whether or not they have ADHD.

The numerator breaks it down into two parts. It's measuring children between

the ages of six and 12, newly prescribed ADHD medication with a follow-up visit by the prescribing practitioner within 30 days. And part two -- this is another study that's multi-faceted. Part two is the continuation phase where the kids have two subsequent visits in months two through nine. So number one, follow-up within 30 days; number two, two additional follow-ups months two through nine. And the denominator statement is all kids getting prescribed ADHD medication.

It's a process measure, and as we go through this, I'll -- I'm a clinician, I'm going to primarily give my clinical impression. I am not a statistician, a researcher and maybe a little bit of administrative work, but that's not where my primary focus is.

In terms of importance to measure, some of the -- review some of the comments from our group. Obviously a lot of kids are getting prescribed ADHD medications, rightly

or wrongly, and it is important to establish follow-up. Some of the questions pertaining to the evidence from my perspective, breaking it down into the -- how often do these follow-ups occur, I was questioning the evidence to support the one-month mark for follow-up visits and should that necessarily be used as a quality measure as opposed to six weeks, two months? And I had trouble finding the evidence to say that one -- I mean, I think we all agree that the follow-up is important and necessary. But in terms of evidence to say how frequently should that follow-up occur, I think that's lacking.

And one of my concerns with this measure is it takes away from services to meet the individual's needs, patient and family.

In terms of performance gap, I think we can all agree that there is a performance gap. I don't know if there is anything else to say about that in terms of establishing that follow-ups should be arranged and there's

different docs, different providers do different things. ADHD is common and it is important to treat.

Some of the -- in terms of importance, for follow-up, some of the comments include with regular follow-up visits, intuitively speaking, it should enhance good medication compliance, engage the patient and the families in the treatment process and adherence.

Should we stop there or should we continue through the whole -- okay.

CO-CHAIR BRISS: You can stop there.

Larry?

MEMBER MILLER: Thank you. I think the committee agreed that this was an important area to look at and I'm certainly glad that NCQA is looking at outcome measures because I think that's one of the things we struggled with that we really wanted to see this as an outcome. And the other thing, as David mentioned, this has nothing to do with

children who were diagnosed with ADHD, they're just given medication for this -- given medication.

Some of the comments that -- just to follow up. There was some concern that there was little improvement using this indicator and that less adherent children were to fall out of the indicator based on the prescribed schedule and there was some concern because those are the kids who may need the follow-up more than anybody else.

I think we all thought this was an important measure to look at and it certainly was a high priority given the use of stimulants and the way that this diagnosis can be thrown around. So those are my comments.

MS. HUDSON SCHOLLE: So this does focus on children who are using the ADHD medication. That's because our early testing work showed that the medications were prescribed when the medications were prescribed. We might not see a diagnosis of

ADHD on the claim but when we went back and looked at the medical record we did. And so we would be under-counting a lot of children -- if we required the diagnosis in the medication, we'd be missing more children.

This measure focuses on children with a new episode and this gets at your issue of, you know, how frequently should these follow-up visits be? The guidance generally doesn't give us a whole lot of -- the guidelines and testing generally don't tell us exactly what the timeframe is. They say it should be addressed based on the needs of the children. But I think what our review panels, our advisory panels said, if this is a new episode then a visit within -- a new episode of treatment, a visit within 30 days is a reasonable expectation for that beginning, right?

So as we had to specify this measure we had to focus on places where we were pretty confident that that was the right thing to do.

So new prescription, you know, without a previous history over the past four months, then you should be checking to see is the child responding or having any side effects? So that's where the one month came from.

The second part of the -- the second indicator looks at this longer-term follow-up approach. You know, like are you actually checking to see how kids are doing? And for that it's -- you pointed out that the decision that this measure made was to say, well, we're going to focus on children who remain on the medication. Now of course, children with ADHD may get that one prescription and not ever get it -- you know, not refill it, they drop out of our denominator. And we don't know whether that's because their symptoms magically resolved, the medicine worked, whatever, they were getting behavioral therapy and that worked. We don't know about that, we're only -- but through the claims data the only thing we can really assess is children are staying

on the medication, therefore someone should be monitoring them while they're on the medication.

And so again, you know, it's the limits of what we can measure from claims data easily. It is a limited measure and we are looking at ways to look at outcomes but we're finding that we're -- there we absolutely have no data. We actually have developed a measure for ONC through -- and CMS, and we're actually -- while we have a measure we can't find any place that can actually test it because they're not -- they don't have a standardized approach to using a patient-reported or family-reported outcome measure over time that would allow us to actually see whether kids are improving. So we're working on that in terms of it's a demonstration rather than a testing point.

So while we recognize that process measures based on claims data are limited, they have their limitations, they're actually

what we can do today.

You mentioned also that this -- the measure is not improving. Okay, this is -- you know, over time what we've seen in a number of our measures that look at behavioral health conditions, so it's not just this measure, it's other measures that are -- address behavioral health issues, we're not seeing improvement. So it's -- and that has to do with how the measures are used and who's paying attention to them.

And so what we've done within our own programs where we have some opportunity to influence this is that, one concern we have is that managed behavioral health organizations or the behavioral health side and the general medical side may not talk to each other enough. And one thing that we see is that --

MEMBER ZIMA: They might.

MS. HUDSON SCHOLLE: They might not. But one thing that we see is that plans that are not responsible for the behavioral health

benefit of their members do not report this measure, right? To be responsible, the plan has to have both the responsibility for general medical and pharmacy and behavioral health, right, so that we can actually measure what's happening.

And so what we have done is, within our -- but we do also credit managed behavioral health organizations and we created a new expectation for those organizations that they begin to track this measure and other measures. There's a suite of measures for them to track because we're trying to get the managed behavioral health organizations to be looking at the same quality metrics that the health plans are looking at.

Neither of them has all the information they need, and so they have to work together, and so that's why -- that's our theory about that, is trying to encourage that collaboration. It's a message I would also give to states that are carving out behavioral

health and employers who carve it out, that they create a boundary unless they can force the sharing of data.

So those are some of the issues that we think contribute to the lack of improvement, but I think it's not that this is a sound measure, it's that there are other things in the environment that are making this and all the -- making it harder for us to improve on a number of behavioral health measures.

CO-CHAIR BRISS: So there are a number of cards up. Why don't we start -- let's start with down at -- the mic down at the end and we'll just work around the table.

MEMBER ZIMA: Okay. And I was reviewing number two on workgroup one. And I think that, you know, it was really -- again, so no change in improvement but just to point out that on page 18, NCQA reports that, quote, "over the past three years this measure has shown improvement." So I -- there was a

1 little bit of a discrepancy, I think.
2 MS. HUDSON SCHOLLE: Over time it's
3 shown -- sorry.
4 Over time we've seen improvement.
5 Over the past couple years, though, it's been
6 pretty steady.
7 MEMBER ZIMA: Yes, it looks quite
8 stable.
9 The other issue is, in the indicator
10 itself it says "med adherence," but I think
11 you're really measuring med prescription
12 persistence with the assumption that it's
13 adherence. But I think to be more specific it
14 should be "med prescription persistence."
15 And the other issue, just to echo
16 some of the discussion, Sarah, that you're
17 talking about, limitations of health plan data
18 is, of course, and it's admitted in the
19 measure, is that you can't stratify by race
20 and ethnicity. And you know, at the same time
21 in the measure application there is a good job
22 as far as reviewing the literature about the

disparities we know about, kids and ADHD.

The other issue that hadn't been commented on yet is that the rationale for using the telephone visit to be counted as one of the follow-up visits, are we talking about specification yet? Or --

CO-CHAIR BRISS: No.

MEMBER ZIMA: Okay. So I'll hold down that comment.

CO-CHAIR BRISS: So before we work our way around the table, Caroline has a question.

MEMBER DOEBBELING: Thank you.

Hello?

CO-CHAIR BRISS: Hello. Yes, Caroline, please go ahead.

MEMBER DOEBBELING: Thank you. I wanted to make sure you all could hear me.

I had a concern about the measure in response to NCQA's statement about the split between behavioral health plans and medical health plans, so to speak, if the behavioral

health is carved out. Given the prevalence of ADHD and the treatment of ADHD by primary care providers and pediatricians, I'm not necessarily so concerned about the behavioral health carve-out with regard to explaining why this measure hasn't improved given that the bulk of prescriptions for ADHD occur in the primary care setting, and the health plans would have those data to report. So I'm not sure that that really is something that makes sense here.

The other question I have, and this might -- it is a question and it might be better for when we talk about the specifications. But in the numerator statement, where does ADHD have to be on the list of diagnoses that go onto a claim? Can it be anywhere in that list or does it have to be the primary diagnosis? Because often ADHD will be addressed in the context of other issues that the child is coming in to see the primary care provider for. That is a question

and I'm not able to tease that out from the description in the numerator.

MS. HUDSON SCHOLLE: It's not actually required at all. It's the medication alone that gets the child in the denominator. So we don't look for the diagnosis, we just look for the use --

MS. DOEBBELING: I -- okay. So then any visit for any reason by any provider in the 30 days or the nine months counts by any prescribing provider?

MS. HUDSON SCHOLLE: Okay. So we're not looking for the diagnosis for the -- it's the medication management is what we're looking for in the visit, not the diagnosis.

CO-CHAIR BRISS: So with that we'll work our way around the table. So starting with Mike.

MEMBER LARDIERI: Yeah, I guess my question is around the same area and I may need clarification.

So it's the follow-up visit with --

is with any prescriber? It says any practitioner. So have you found that the -- especially with people churning in and out of plans, it might start with a psychiatrist and then go to a medical provider and not really have any follow-up because they didn't do any medication reconciliation and that kind of stuff. So how does that play out in the measure?

MS. HUDSON SCHOLLE: Actually, I think we made a mistake. It looks like it's a principal diagnosis, a principal mental health diagnosis. We're just not trying to be specific about the ADHD given that the -- that children, this can be combined.

We say -- let's see -- the denominator details -- I'm looking at the numerator details. Hold on.

(Pause.)

MS. HUDSON SCHOLLE: And do you have the value sets?

MS. FRANKLIN: We'll get back to

you. I want to hear --

CO-CHAIR BRISS: Yeah, let's continue the discussion and you can come back when you've found it.

And were you done, Mike?

MEMBER LARDIERI: Yes.

CO-CHAIR BRISS: Vanita.

MEMBER PINDOLIA: So this question comes from -- so I work with a health plan with multiple providers and then I work with our provider and an ACO for multiple health plans. And trying to improve this measure has been a great struggle.

And one of the comments I have is, you know, to understand really the data of having that follow-up visit within 30 days being so critical to demonstrate an improvement in medication management for these drugs. That is very difficult to get parents to come in for a second copay within 30 days. It's just -- and that's with providers' inputs, from multiple stakeholders on that

side with different health plans. So I think that might be why that number, if you look, is even lower than your long-term, which usually is the other way around for other measures.

The second part is understanding the complexity of when these patients are going through to school and if they were diagnosed in March but then in summer they take off the drug and then they resume in September, does that continuation and breakup count as the three-month has to occur in July or August when they're technically off the pill? But if you do it continuous throughout the year, in September they got their fill again. So the health plans and the physicians feel they get dinked but they didn't really need to see that patient in July. That decision was done at the end of the school year.

MS. HUDSON SCHOLLE: So the prescription carries over the summertime but they're not actually taking it even though they have the prescription? But would they

get a new prescription filled? Because we're looking at the prescription fill has to cover the 210 days.

MEMBER PINDOLIA: So the summer is only the three months. So within that 210 days there will be a prescription at the start of that phase, and the end of the phase, so they get included into the denominator. However, we have a very difficult time getting a office copay within 30 days. But then looking at that three-month or that six-month, if that lands in that summer period there's no parent or doctor that feels it's a necessity to even have that visit. So that makes it really difficult to understand where we're missing the holes. Is it the three -- is it because of the summer months? So I have some concerns about the measure of them, how much it's really improving quality.

MS. HUDSON SCHOLLE: Okay.

MEMBER ROBINSON BEALE: I'm going to comment as a system administrator and from a

health plan perspective. While I do understand the complication and I do think the issue of the summertime medication-free issue really needs to be taken in consideration, it's complicated because in different states they have different types of rotation for children being off. So I'm not sure if that's doable or feasible, but I agree that it's a very important piece.

I look at this measure not as an improvement measure, I look at it as two things. One is an accountability measure and a measure of medication safety. What you're actually measuring is whether or not someone who's prescribed a stimulant for a child has actually followed them up. And when I look at it from that perspective, that's a very important issue that I think deals with patient safety but also the accountability of a provider who's actually prescribed a drug for a child.

So when I look at it that way, it's

a very important significant measure because we know of so many children who are placed on medications, don't need to be on medications and are never followed up.

The issue of the timeframe of the 30 days, I think we're unfortunately trapped with the guidelines which is no better than offering the 30 days. And I respect that NCQA does abide by the guidelines that are put out there. It does raise a question as to whether or not the guidelines have actually looked at the incidence of side effects and when they occur and when they need to be reviewed. It might be helpful to take a look at the fallout in your second measurement in terms of office visit to determine whether or not you're getting a fallout because, one, the population may have had side effects but there's been no -- there's 30 days before they're followed up and so the parent and the child just stop and they just don't come in. So that's a complicating factor that could, in some way,

defeat the purpose of this being a medication safety and an accountability measure.

The other thing that I wanted to say is that this is also an indicator measure. It's an indicator of the strength of the system to be able to foster that kind of follow-up. I hear what you're saying about the copayments. Copayments are a big problem across the board and they're particularly a problem for behavioral health in terms of the copayments being, despite parity, still in some ways causing more problems in terms of higher dollar amounts. And particularly for something that may require -- in most states they do require that the prescription be rewritten in 30 days, and so that's a consistent schedule that one has to shell out dollars for the copayment. But to me, that's a plan issue and that's not an issue of measuring the accountability of a prescriber and the medication safety.

CO-CHAIR BRISS: There are lots of

cards up. And so I encourage us to be as efficient as possible with this time.

MEMBER EINZIG: So I'll try and be brief.

I think this is similar to the previous measure where you're putting multiple factors into one. Reasonable to have regular follow-up visits to ensure quality and safety but the question of that is that 30-day mark for me. And the lack of evidence to say that that necessarily improves outcome, I think that's missing.

I reviewed the AAP practice guidelines and I didn't see anything in there for one-month follow-up. In the AACAP, practice parameter, American Academy of Child and Adolescent Psychiatry. One of the last statements in the summary, I'll just repeat it word for word, "although this parameter does not seek to set a formula for the method of follow-up, significant contact with the clinician should typically occur two to four

times per year in cases of uncomplicated ADHD and up to weekly sessions at times of severe dysfunction or complications from treatments. Nothing in this parameter should be construed as justification for limiting clinician contact by third-party payers or for regarding more limited contact by the clinician as substandard when clinical evidence documents that the patient is functioning well."

So in other words, perhaps regular follow-up visits are a good indicator of quality but I think it's inappropriate to say that if a patient doesn't follow up with a prescribing provider in the first 30 days, that that necessarily assumes that is poor quality.

The other comment I would have is in this world that we're moving more towards collaborative care models or shared care models, if the psychologist is down the hall there's going to be frequent dialog with the prescriber and the psychologist. Or if a

psychiatrist does the initial prescription and then the kid with cancer has a follow-up with a hematologist/oncologist a few weeks down the road, gets their vital signs and no cardiac concerns at that visit, does that necessitate a follow-up with the psychiatrist to make sure that things are safe?

And one final comment, I'm trying to be brief, nobody knows the patient and families better than the clinician, it's all about the relationship. So for -- just to make it real-world, so I see families with several kids with autism, not necessarily ADHD but prescribing similar medications to help with their impulsive reactivity. They have developmental issues and they don't like change and don't like being taken out of school. And so -- and you know this family well because you see multiple kids in their family. That might be a specific example where follow-up might not be clinically indicated in the first month.

I can go on but I'll stop.

CO-CHAIR BRISS: Thank you.

So again, either lots of -- to the extent that there are concerns about this measure, they seem to center around the specifics of the follow-up that's being recommended, that it's a -- I'd encourage us to -- there's still a lot of cards up there. I'd encourage us to be brief, to continue to try to be brief and to try to raise issues if they're new issues.

MEMBER CHALK: I want to echo what Ron just said about this measure. I think as it's identified as a follow-up, a follow-up measure, I think that that's a misnomer. I think it's a medication management measure. I think it's a safety measure. And I think it's -- I think that makes every difference in the world in terms of how we think about it here. And I'll leave it at that.

MEMBER TRANGLE: I think my comments were actually already covered. They were

about behavioral health home and how do you sort of factor in efficiencies with still accountability. I would like at least to ask the developers to think about how can you incorporate internet or video, kind of working with a care manager in the clinic and not always seeing a prescriber every visit so that you can capture.

MEMBER SUSMAN: And from the primary care side as we're moving to medical homes that are fully integrated that are using a diversity of personnel and staffing, having a visit frankly in this condition, particularly, has pretty little purpose. I mean, you can go through a symptom checklist and ask about side effects from a nurse or someone else in the office, do it in a much, I think, more futuristic approach to these chronic conditions. So I think the underlying basis for this measure is more historical rather than forward thinking. And better systems could be penalized by integrating behavioral

healthcare more broadly.

MEMBER MARK: I had a question about the specification of the medications defined as ADHD meds on page 12. Like clonidine, is that a indicated use for ADHD or is that off-label?

MS. HUDSON SCHOLLE: Yes.

MEMBER MARK: It's -- I'm not a clinician, I just --

MEMBER EINZIG: May I? Generic clonidine technically is off-label. The new version of it, Kapvay, which is the extended release is FDA approved. The reason for that is historical, that clonidine is generic and nobody was going to make money off of doing studies on it. But it is commonly used as an ADHD medicine.

CO-CHAIR BRISS: Mike, do you still -- do you want to have a last word?

MEMBER LARDIERI: Ask Caroline.

CO-CHAIR BRISS: Oh, and Caroline. I'll give Caroline the last word. I'm sorry,

David.

MEMBER DOEBBELING: Thanks.

I'm still not clear again about the specifications because as I read them it is for any practitioner with prescribing authority within 30 days after the earliest prescription of dispensing the new ADHD medication which means that it's not necessarily, I think, the prescriber who started the medication doing that follow-up visit. It would appear that it could be any prescriber in any of the intensive outpatient settings or other outpatient settings. So I'm not sure that it even gets to any of the comments about this is a safety measure or this is an adherence measure because it may not even be the same person doing the follow-up or the same clinic doing the follow-up.

MS. HUDSON SCHOLLE: And that is -- we did confirm, that is true. It is the prescriber. Again, we're limited to what we can capture from the claims data and being

able to limit it back to the original prescriber in the claims. It turns out to be tricky, who'da thunk. But so I did want to point that out that and I appreciate the discussion about this is an outdated measure and we should be thinking about when other kinds of visits should count and other kinds of team-based approaches to care and how we should do that.

I'd point out that this initial -- for this measure, if plans are able to track those kinds of visits, track telephone calls, again they have to be with the prescriber. But if they had those data systems that allowed that then a telephone contact with a prescriber will count for the maintenance phase of the measure. But we still have that 30-day, it needs to be a face-to-face for that initiation. Again, you know, the research -- that this measure was developed based on the research that weekly visits and it was just standard in the research trials that showed

that ADHD medications were effective, looked at follow-up every week.

Now since that time, the guidelines have changed a little bit and have loosened, but we're still looking at children that are getting their -- you know, a new episode and so that's why we've kept that 30-day phase and it is face-to-face.

CO-CHAIR BRISS: So I want to get to the end of this discussion soon. I know I've let this go for a while because it's -- it seems to me that the importance of the measure sort of hinges on the -- whether the details of this specifically describe follow-up is going to result in better outcomes for these kids. But I do want to get to the end of this. So please, if you have new things that haven't been touched on already, quickly make those points. And otherwise, I'm going to -- Bonnie?

MEMBER ZIMA: Peter, you -- talking about specifications reliability now, which is

the evidence.

CO-CHAIR BRISS: No, we're talking about importance to measure.

MEMBER ZIMA: Okay.

MEMBER PINDOLIA: So for a measure that's been out already for three or six years and was developed in one way, is this a time that we can make a recommendation because of how practice has changed? And is that part of that process or no, we can't?

CO-CHAIR BRISS: You can. You're voting on the importance to measure of this measure as of today. It doesn't have to influence you if it's been out before.

Harold? My co-chair isn't helping me here.

CO-CHAIR PINCUS: Yeah. I just want to -- I want to just -- because I think I'm -- I'm worried about our discussion of importance affecting all of the future discussions of importance. And I think just we need to separate out all the other criteria from

importance. And it's the importance of measuring this concept so that if -- is it worth pursuing? Is it important enough to pursue a measure on this concept, you know, beyond this specific measure? But the concept of assessing, you know, whatever this measure is intended to assess as a concept. You know, not the specifications.

So if all the specifications were perfect, would it still be important to look at this? I think that otherwise, when we discuss importance we're discussing everything. It's like, oh, you can't capture it, it's impractical to capture -- you know, there's no evidence about this. And otherwise it becomes a gemish. And so that this is -- it sort of separate out this concept of importance for this measurement concept. And without regard to this specific measure but just so we can keep saying as we go through this.

CO-CHAIR BRISS: So with that, I'd

love to take it to a vote. And so importance to measure, please.

MS. BAL: So the first thing we'll be voting on is 18, evidence. And then the choices are one high, two moderate, three low, four insufficient evidence, five insufficient evidence with exception. And you can vote now.

(Brief pause.)

MEMBER SUSMAN: By the way, what does insufficient evidence with exception mean? I'm not voting that way but I'm not even sure I know what that means.

MS. FRANKLIN: That would -- this is Angela. That would mean that you found that the evidence presented was not support -- was not supportive of the measure but you felt like it was important enough to make an exception. And if you look at your algorithm you would have to go through an analysis starting in box ten of whether there were other measures out there that could be --

(Laughter.)

MS. FRANKLIN: I know, we're trying to be very formal about this. So you would want to look at this algorithm steps to make that exception.

MS. BAL: Okay. So the results are for evidence of 0108, it's high seven, moderate nine, low five, insufficient evidence one, insufficient evidence with exception zero. And we'll move forward with this measure.

CO-CHAIR BRISS: So with that we'll move to reliability and validity.

MS. BAL: Actually we'll move to gap. Yes. So we'll move to gap, and again the options are one high, two moderate, three low, four insufficient, and voting is now open.

(Brief pause.)

MS. BAL: Okay. So for performance gap for 0108 the results are high nine, moderate eleven, low one, insufficient one.

And we'll move forward to high priority. And again the options are one high, two moderate, three low, fourth insufficient. And we can start voting now.

(Brief pause.)

MS. BAL: And also we're looking for 22 votes for this round, so everyone's aware.

(Brief pause.)

MS. BAL: All the votes are in, we're just waiting for Caroline now.

(Brief pause.)

MS. BAL: Okay. The results for high priority for 0108 is high twelve, moderate seven, low three, insufficient zero. And we'll move forward and now we can speak.

CO-CHAIR BRISS: And now reliability and validity. And we did tee up many of these issues in the first discussion. So David, will you tee this up first?

MEMBER EINZIG: Yes. So again I'm not a statistician, I'll do the best I can to talk about reliability. So this is a measure

from the healthcare plan level, healthcare claims looking at prescriptions being filled. As a group I think the majority of us did find, you know, it was reliable, other than the -- perhaps the dropout rate, the summer issues, kids being off medications in summers and whatnot. I'm not sure if any other folks from the committee have other things to say about reliability then we can move on to validity.

MEMBER ZIMA: So I was Reviewer 2 on this and as far as some -- a few issues, one is that during the CNM phase in the specifications, a telephone visit is -- can be counted as one of the follow-up visits. And when you look at sort of the fine print in this, the rationale is that within the Academy guidelines the method of contact is not stated and that's what's stated on page 25.

However you know, I think you had made a comment earlier, Dr. Susman, about that, you know, maybe it was okay to just have

somebody just get parent rating scales and not have that face-to-face. But I would argue, actually in clinical practice you actually have to lay eyes on the child and you should interview him to assess med safety and efficacy because there are times where parents actually aren't good informants. We're working with very distressed parents and sometimes you really do need to see the child.

The other issue is that the implication of accepting a telephone visit is that a health plan may have an acceptable pass rate but children with med side effects or ineffective medication treatment could be under-detected with that telephone visit. So I think that frankly when I looked at the Academy guidelines, they didn't specify face-to-face but I think the assumption on the folks that created the practice parameter was that it was a face-to-face contact.

So, and then the issue of reliability, I think -- and you see these in

a number of the NCQA measures, and that's because it is a claims data so they're using a beta binomial method which is basically detecting the extent of variance due to a health plan compared to measurement error. And so what you'll see is that, you know, when they did the cross check between commercial and Medicaid, the average binomial wheel went up under Medicaid. But I was wondering frankly if this is because SCS was held constant?

And then I think, you know, throughout the day we'll be talking about whether the approach of face validity based on an advisory panel is adequate, and then also you'll see in the construct validity where they looked at the correlation of the measure with access to primary care. And that made sense. And actually I was kind of surprised because I thought, since access was required in the measure, you would have had an even higher correlation coefficient but it was only

point-four. So I think those are my comments on specification reliability.

CO-CHAIR BRISS: Sarah, thank you. And again we've teed up several of the reliability and validity issues in the previous discussion. Are there other things that haven't been raised that the committee wants to hash out?

Yes?

MEMBER GOLDSTEIN GRUMET: So I just want to reiterate what you said because I think my concern is, by doing a review of the chart or the records, you don't actually know what they're assessing and asking. And so it's a concern that you're actually assessing the quality of that interaction and whether that actually means the child's getting better because there's inconsistencies across providers. So to me, a better measure is somehow reviewing the chart to see what exactly was asked or some kind of -- and I mean, this is a different question, but

ensuring some kind of a measure that actually provides the list of questions or some questionnaire. I just get concerned that this doesn't -- this shows that they're taking their medication potentially but it doesn't actually show from provider to provider what that -- what the interaction is like.

MEMBER SUSMAN: I'm just getting a little concerned that we're holding ourselves or the developers and measurers to a very, very high standard. While I am the first to have shot holes in some of these measures I also think that at the end of the day we should ask, are these measures valid and reliable enough to move here forward in a positive direction. And you know, in some cases I think we may not have ties directly to the ultimate outcomes or all the nuances that we might wish. On the other hand, if it starts us down a causal pathway of improvement I think there's a lot of positives to that.

And I've looked at depression in our

sort of movement in the depression world from sort of measuring did we ask about it, did we measure it? Now we're finally getting to, hopefully, outcomes and improvement. I think we're probably in the same way in a lot of behavioral health fields moving from a incidence prevalence measures to looking at improvement and ultimately looking at outcomes that are long term.

MEMBER MARK: For me, I think part of the calculus here again is what's the harm of, you know, maybe erring on the side of something that's not perfectly reliable in validity. In contrast to the other measure I don't feel that there's as much harm here if we have not perfect reliability and validity. And maybe people want to counter that perception.

MEMBER PATING: I'm really just trying to look at the evidence. And with regards to the one-month indicator and the reliability of that, I mean, it's fine to have

a consensus panel driving, you know, some practices. But what I'm very, very concerned is that these practice -- these measures are getting so specific and they've really substituted in many systems for clinical judgment that is not just like a quality outcome, you're driving real practice such that these measures actually supplant the -- what is the prescribed care paths or even, you know, what an individual clinician wants to do. Because there's a whole systemic weight and quality measures and performance outcomes and incentive bonuses that are just -- that are tied to these measures and they're not insignificant. So it would be fine to say this is a general direction we want to go but now we're saying everybody across the nation has to have a follow-up in 30 days and there's no evidence except a consensus panel that was basing it on we don't know what.

So I just really want to be -- I'm actually going the opposite direction, Jeff,

as being hyper cautious because I'm seeing the multiplying effect of this at the grassroots level as very profound. Clinicians are just under -- this is affecting their pay if they don't do this, and they're going to do it because, you know, hook or crook, regardless of whether the patient wants to come in at 30 days or it's clinically indicated, whether there's an extra cost, that's how -- that's the strength that these indicators are gathering.

CO-CHAIR BRISS: Necessarily quickly, I'd like to move down this row and up the other one and then let's -- so let's try and be efficient and get to -- Larry?

MEMBER MILLER: Yeah, I think we're having the discussion looking at this as an outcome measure and it's not, it's a process measure. It simply raises awareness that kids who are on these medications need to be followed up. And I think that's a very good beginning, I think that this is a very

important issue to look at and I'd just like to keep us in mind of that.

CO-CHAIR BRISS: David?

MEMBER EINZIG: So I think nobody's arguing that follow-up is important but it's the evidence behind -- we're asking to be -- speak to the quality of two things. Follow-up within one month and a couple additional follow-ups in the subsequent nine months. And you know, same thing as the first measure, it's just that question of is it, you know, one vote for all measures when half of it might not be applicable.

And also just kind of going back, it wasn't too long ago that with anti-depressant medications the recommendation was we follow up with the prescribing provider within the first month and then every two weeks in the second month and that backfired. You know, people stopped prescribing anti-depressions and suicide rates went up. You know, there may or may not be a direct correlation.

So I really want to stand strong on, you know, is there evidence to say the 30-day mark, is there sufficient evidence with that?

MEMBER ROBINSON BEALE: Yeah, I do think all the things that have been said are very, very important. I do think it brings for NQF to have an opportunity to compile a list of questions that need to be answered by those who are developing guidelines in terms of really the evidence as it relates to the timeframe for certain things.

I realize NCQA uses the guidelines and tries to make the best out of the guidelines but let's go back to where the problem may be and that's the lack of specificity. And particularly with behavioral health, they are very non-specific. And so I do think there's a very strong message that needs to go back to say there are specific things that are needed in order to really drive good measurement, drive good clinical judgment. And I think without that

specificity we're continuously driving the system in a very non-specific and invalid way.

CO-CHAIR BRISS: Bonnie?

MEMBER ZIMA: Yeah, I just want to get back to the issue of do no harm. And I think that I continue to be concerned particularly with just a telephone visit counted as one of those follow-up visits and having only two follow-up visits in nine months with a child on a Class II medication, some of the harm is that you can continue a child on a medication that's ineffective or you aren't detecting med side effects.

CO-CHAIR BRISS: And the last word.

MEMBER MAZON JEFFERS: The last word, huh? I just wanted to go back to what David pointed out. And I have no doubt that increasingly quality measures are being used at the grassroot levels to set payment and performance. But if you think about what quality measures are really supposed to do, this measure in particular as it's looking at

system-wide performance, it actually sounds like it is already doing some of the things that it set out to do. So the observations that you made that the copay is a hindrance to a 30-day follow-up visit, these are really important systems that, by applying the measure, you have learned as a system.

And similarly the issue of the summer experience is also something that the quality measure has diagnosed, if you will, about the system of care that needs attention. So if we think about the purpose of quality improvement measures, it is driving the quality of care of the system when it's used at a system-wide and not necessarily, of course, there will be individuals for whom a 30-day follow-up visit would not be clinically indicated. But my guess is they might be in the minority and outliers rather than the main -- the mean.

DR. PIERCE: Hi, it's Karen Pierce and I didn't know I'm in the middle but I

didn't know my line wasn't on. I've been listening.

I'm a child psychiatrist and let me tell you my concerns about -- behavioral health measures are really very hard to --

MS. FRANKLIN: Excuse me --

DR. PIERCE: Yeah?

MS. FRANKLIN: Who's speaking?

DR. PIERCE: Karen Pierce, I'm a child psychiatrist.

MS. FRANKLIN: Okay, thank you.

DR. PIERCE: I helped develop these measurements a bit ago.

I want to add that developing behavioral health measurements is -- we're nowhere near the level that we are in the medical level.

MS. FRANKLIN: Excuse me, ma'am, we're not having public comment at this time.

DR. PIERCE: I understand. But the issue you have is about 30 days --

MS. FRANKLIN: Kathy, could you

please say -- let us know when we have a comment?

Okay, I'm sorry, this is not the measure that you were a developer for. I'm sorry.

DR. PIERCE: Okay.

MS. FRANKLIN: We're looking right now at the NCQA measure.

DR. PIERCE: Right. Okay.

MS. FRANKLIN: And we're getting ready to vote as a committee. Thanks.

MS. BAL: All right. So we'll be voting for reliability. The options are one high, two moderate, three low, four insufficient and voting is now open.

(Brief pause.)

MS. BAL: Okay. So we have for reliability for 0108, high one, moderate fourteen, low four, insufficient three. And we can move forward with this measure to validity.

And we're ready to vote for this.

So the voting for validity is now open for 0108. Again the options are one high, two moderate, three low, four insufficient.

(Brief pause.)

MS. BAL: Okay. So the results for validity for 0108 is high two, moderate fourteen, low four, insufficient three, and we can move forward with this measure to feasibility.

(Brief pause.)

MEMBER EINZIG: Okay. Feasibility, so data collection is from the healthcare plans. I don't think there were any issues -- I'll keep it brief, I don't think there were any issues with the feasibility.

CO-CHAIR BRISS: That was mercifully brief. Thank you.

(Laughter.)

CO-CHAIR BRISS: Bonnie, anything to add?

MEMBER ZIMA: No comment.

CO-CHAIR BRISS: So any -- we have

had some feasibility discussion already. Anybody want to make further comments that haven't already been made?

(No response.)

CO-CHAIR BRISS: Maybe it's a positive thing to be moving toward break. So let's move to voting.

MS. BAL: Okay. The voting for feasibility is now open. Again, the options are one high, two moderate, three low, four insufficient.

(Brief pause.)

MS. BAL: Okay. The results for feasibility 0108 is high eight, moderate fourteen, low one, insufficient zero, and we'll move forward to usability.

CO-CHAIR BRISS: So I think the last conversation -- we've already had a lot of conversation on usability so just to sort of highlight some of the issues and concerns. Some of David's comments about using this as a quality measure, data collection when there

may be lack of evidence to say that it truly does improve quality of care, lack of evidence behind the specific timelines of when follow-ups should occur. So again I think everybody agrees follow-up, regular follow-up is important and necessary, it improves quality. I think the issue has to do with standardizing care to one-size-fits-all model versus meeting the individual's needs for that particular child and family.

Bonnie, do you have anything to add?

MEMBER ZIMA: No, no additional comments.

MEMBER MILLER: I'm sorry, I just want to quote one of the comments, I think it's a useful comment.

"Although this parameter does not seek to set a formula for the method of follow-up, significant contact with a clinician should typically occur two to four times per year in cases of uncomplicated ADHD and weekly sessions at times of severe

dysfunction or complications. Nothing in this parameter should be misconstrued as a justification for limiting clinician contact by third-party payers or for regarding more limited contact by the clinician as substandard when clinical evidence documents that the patient is functioning well."

CO-CHAIR BRISS: So any further comments?

(No response.)

CO-CHAIR BRISS: Hearing none, let's move to a vote.

MS. BAL: Okay. The vote for usability and use is now open. The options are one high, two moderate, three low, four insufficient.

(Brief pause.)

MS. BAL: Okay. The results are, for usability and use for 0108 is high four, moderate thirteen, low six, insufficient zero, and we will move forward.

Do you want to go ahead straight

into the vote or do you want to have discussion first?

CO-CHAIR BRISS: So any further discussion before we do an overall suitability vote?

(No response.)

CO-CHAIR BRISS: Hearing none.

MS. BAL: Okay. voting is now over -- not over -- open for overall suitability. The options are one, yes; two, no. Again, the options are one, yes; two, no.

(Brief pause.)

MS. BAL: If everybody could just vote one more time? We're missing one vote. The number is now 23. Okay, perfect. Thank you.

All right. So for overall suitability for 0108, we have yes, seventeen; no, six. So this measure is recommended.

CO-CHAIR BRISS: Thank you all.

We're slightly behind. We'll take a 10-minute instead of a 15-minute break. And

I'd like to encourage us to be back seated and be ready to go in ten.

(Whereupon, the above-entitled matter went off the record at 11:21 a.m. and resumed at 11:33 a.m.)

CO-CHAIR BRISS: So please be seated. And while people are seating themselves maybe -- we've had several new people join us and so maybe we can do some additional introductions.

CO-CHAIR PINCUS: Let's everybody get seated.

So there's a couple of people that we wanted to ask to introduce themselves. Les, if you could introduce yourself?

MEMBER ZUN: Good morning. Do you want me to give any background?

CO-CHAIR PINCUS: Please give just your background.

MEMBER ZUN: I'm the token emergency physician in the group, I believe, but I am kind of an oddball in that I actually am very

interested in behavioral emergencies. I've been doing research in behavioral emergencies, wrote a textbook on behavioral emergencies, do a conference every year on -- or I do a conference on behavioral emergencies every year. And I'm the president elect of the American Association for Emergency Psychiatry and I also sit on the Board of the American Academy of Emergency Medicine, and we're a new organization member. Thank you.

CO-CHAIR PINCUS: Is somebody on the phone?

(No response.)

CO-CHAIR BRISS: Helen Burstin has joined us.

DR. BURSTIN: Yes.

CO-CHAIR BRISS: Helen, would you run to the table and introduce, please?

DR. BURSTIN: Apologies for being late.

I'm Helen Burstin, I'm the Chief Scientific Officer here at NQF. Thank you

all, for many of you returning again to this committee and for some new faces as well. Thank you to join us in this very, very important ask. Thanks.

CO-CHAIR PINCUS: Committee member on the phone?

MEMBER DOEBBELING: Yes, I'm still on.

CO-CHAIR PINCUS: Everyone wonders if you got a chance to introduce yourself.

MEMBER DOEBBELING: I did earlier, thank you, Harold.

CO-CHAIR PINCUS: Okay. Good.

Now we're going to address measure 1365 and it's a -- the measure developer is the AMAPCPI group. And so Kendra, do you want to introduce it?

#1365: Child and Adolescent MDD: Suicide Risk Assessment (PCPI)

MS. TIERNEY: Hi, everybody, and thank you for the opportunity to be here and present to you. I'm Sam Tierney, I'm with the

AMAPCPI in our measure development group and I'm here with my colleague Kenra Hanley who's in our measures classifications group. We have several people on the phone as well as Dr. Pierce who is a child and adolescent psychiatrist who helped us in the development of the measure.

So just by way of background, I wanted to explain a little bit about how the process that we used to develop the measure, just so you have a sense of the rigor that's involved in our process. The measure was developed in late 2007 as part of a set of measures for child and adolescent MDD. Together with the Child Psychiatric Association and the American Academy of Child and Adolescent Psychiatry the AMA convened physician consortium for performance improvement or PCPI, it's a mouthful. We formed a workgroup to identify and define quality measures toward improving outcomes for patients with child and adolescent major

depressive disorder.

The workgroup included a cross specialty, multi-disciplinary workgroup of key stakeholders representing a variety of disciplines; general and child and adolescent psychiatry, pediatrics, family medicine, internal medicine, emergency medicine, we had some government representatives as well and health policy folks.

The workgroup was charged with developing measures with the strong clinical evidence-base and based on areas that they felt there was a need for performance improvement. The measures were well vetted by a diverse group of stakeholders through a public comment process which was the community enabled comment process and we also submitted them for approval to the PCPI membership which is a broad group of medical and state specialty societies as well as other quality organizations.

So we feel that this measure

addresses a high-impact topic area. We noted some things on our form but if I could just highlight a few of them to start the review of the measure. MDD is a debilitating condition that has been increasingly recognized among youth, particularly adolescents. The prevalence of current or recent depression among children is three percent and among adolescents is six percent. The lifetime prevalence of MDD among adolescents may be as high as 20 percent.

Research has shown that patients with major depressive disorder are at a high risk for suicide attempts and completion. Suicide is the fourth leading cause of death among all youth and young adults between the ages of 10 and 24 years, accounting for 15 percent of all mortality in that age range. We think that suicide risk, and others as well, and guidelines have felt that suicide risk is a critical consideration for this patient population and it's important that it

be assessed at every visit. Suicide risk assessment can help to ensure that early detection of those ideations, referral and treatment for patients at high risk of suicidal behaviors.

We have data in the forms that indicate that there is a significant opportunity for improvement although it might seem like this is a routine part of care. And we believe that the importance of the assessments for suicide risk is really underscored by research that indicates that many individuals who die by suicide have made contact with primary care and mental health care providers recently before their death.

Just a little bit about the measure. The measure has been selected for use in several national programs. It is in the Meaningful Use CHR Incentive Program and also in the Physician Quality Reporting System.

So with that, we look forward to your discussion. Thank you.

CO-CHAIR PINCUS: Bernadette, do you want to summarize the comments specifically with regard to importance? Now I -- just to clarify, I'd like, going forward, that when we talk about importance we don't talk about everything, okay? So really what we're talking about when we're talking about importance, we're talking about the idea of measuring this kind of -- within this kind of focus. Irrespective of this specific measure and its performance characteristics, is it important to measure and assess screening for suicidality among individuals who are depressed? That's the concept, the focus of this measure, and so we're looking at the importance to measuring report about this concept. We're not talking specifically about these measures' specifications, okay?

MS. MELNYK: So related to that concisely, our subgroup believes that this is a very important measure in terms of taking up suicidal risk. The one area of concern by our

workgroup members was the -- that there was a feeling that the evidence that specifically supports the premise that conducting a risk assessment reduces suicide attempts was not presented.

And the second point dealing with importance is the group question going down to the age of six when most children do not develop death as final and irreversible until eight to nine years of age. But otherwise than that, people believed this was an incredibly important measure.

CO-CHAIR PINCUS: Other comments from people around the table?

(No response.)

CO-CHAIR PINCUS: So does the measure developer have any sort of response to the two issues that Bernadette raised?

MS. TIERNEY: Sure. I have a colleague on the phone, Toni? I wonder if you could speak to the two issues that have been raised? And also Dr. Pierce, if you're still

on we'd be happy to have your comments as well.

DR. PIERCE: The issue, they may not have a concept of death but it's still something that can be verbalized and not understood about the finality of death. And so I think it's important that they're surveyed as much as anybody else is surveyed around death and the risk around that. So age, you're right about the understanding but you still can die if you're depressed.

CO-CHAIR PINCUS: I guess just --

MS. KAYE: This is Toni with the PCPI, are you able to hear me?

CO-CHAIR PINCUS: Yes.

MS. KAYE: I guess I'd also like to supplement what Dr. Pierce said regarding the choice of age range. We chose to go down to the age of six in part because the guidelines from the Academy of Child and Adolescent Psychologists, they did specify that the recommendations did apply to both children and

adolescents. And so to kind of support that we did find a 2013 cohort study by Rohde, et al, that showed in their cohort of 815 participants, five percent had their first incidence of MDD between the ages of five and twelve within that younger, and they were scattered somewhat homogeneously throughout that age range, so it wasn't all twelve-year-olds. It did happen. There was incidence at younger ages as well. So we felt that justified having a suicide risk assessment as part of their MDD treatment.

CO-CHAIR PINCUS: What about the second issue that was raised with regard to the, sort of, the proximal relationship between screening and outcomes?

MS. KAYE: Hi, this is Toni again.

So regarding the relation to outcome, I agree it can be tough to directly link screening to suicide rates, either completion or attempts, due to -- even though it's an important area. In order to have

enough incidence of that to statistically significantly link the two can be tricky. However, we did find there was a study from 2010 that showed they had clinics, they did intervention to increase the screening rate and then they looked at the impact on the detection of suicidal ideations and the rates of referral. And both of them, they more than doubled the referral rate, the detection rate of ideations just through increased screening.

And so I would propose that those are equally important outcomes that do have some more evidence and are more proximal to the screening which would be appropriate identification of risk, referral where appropriate or treatment as needed.

MS. TIERNEY: And this is Sam, if I could also add to what Toni said.

There was -- you're probably all maybe familiar with the USPSTF statement on screening for suicide risk in adolescents and adults in primary care, which was certainly a

different population than this measure. But in their review and when they spoke about the effectiveness of early detection in screening they really emphasized the fact that treatment, particularly psychotherapy, has been shown to have an effect, a positive effect on suicide attempts.

So I don't know if the data specifically around a reduction -- screening alone having a reduction of attempts is available but it seems like, based on the USPSTF review that there's clearly a link between the treatment and decreased attempts. And the screening would lead to increased treatment.

CO-CHAIR PINCUS: And I guess just going back to, I guess, something Tami brought up earlier, at an earlier point, is there any evidence to bear on one side or the other with regard to any kind of negative side effects of screening?

DR. PIERCE: I don't know of any

data that says it's negative. I think that's been the myth. I only see that it's been positive.

CO-CHAIR PINCUS: I've seen a couple of fairly rigorous reviews in this area and it's hard to find evidence for asking about suicides or a suicide.

DR. PIERCE: Yeah.

CO-CHAIR PINCUS: Any other comments, questions by the committee?

CO-CHAIR BRISS: I guess the other issue that is -- that this measure raises is the non-standardized assessment. So it's a bit of a checkbox measure. And so --

CO-CHAIR PINCUS: But that's getting into specifications.

CO-CHAIR BRISS: Except that it gets to the question of is this worth measuring, right? And so it's -- the question is sort of are all assessments created equal and is a checkbox measure enough?

DR. PIERCE: My sense is that it's

best to be asking and I think we need to be asking, and I think it's important to be asking about suicide. And so checkbox is what we've got right now and so I think that's where we have to start.

CO-CHAIR PINCUS: Any other comments, questions?

(No response.)

CO-CHAIR PINCUS: Okay. So I think we're ready to vote on importance.

MS. BAL: Okay. Voting is now open for evidence for 1365. And the options are one high, two moderate, three low, four insufficient, five insufficient evidence with exception.

(Brief pause.)

MS. BAL: Okay. The results for evidence for 1365 is high eighteen, moderate seven, low zero, insufficient evidence zero, insufficient evidence with exception zero. And we will move on to the next vote which is performance gap.

MEMBER MELNYK: The committee felt there was a great opportunity because there is a performance gap. There is a lot of variability that exists. And that variability also concerns, there's no standard criteria or a standardized tool for assessment that people use.

CO-CHAIR PINCUS: Okay. Ready to vote?

MS. BAL: Okay. Voting is now open for performance gap. The options are one high, two moderate, three low, four insufficient. And we're looking for 25 votes.

(Brief pause.)

MS. BAL: Okay. The results for performance gap for 1365 are high eighteen, moderate six, low one, insufficient zero. And we'll move forward to priority.

MEMBER MELNYK: So we'll move on to the --

CO-CHAIR PINCUS: Yes, are there any comments about priority?

(No response.)

CO-CHAIR PINCUS: Okay. I think going forward, we'd like to take each of the categories sort of as a group and then -- you know, for discussion and then come back to the voting, you know, boom, boom, boom, okay?

MS. BAL: Okay. Voting is now open for high priority. And the options are one high, two moderate, three low, four insufficient.

(Brief pause.)

MS. BAL: Okay. The results for high priority for 1365 are high twenty-one, moderate four, low zero, insufficient zero, and we will move forward.

CO-CHAIR PINCUS: Great. So now we're going to move to scientific acceptability. And so Bernadette and Bonnie, could you sort of walk through all of the components of scientific and then --

MEMBER MELNYK: Sure. So face validity was assessed with an 18-member expert

panel who responded to one question. Eighty-nine percent of the expert panel agreed or strongly disagreed that the measure can accurately distinguish between good and poor quality. In terms of reliability the measure actually -- psychometric property for reliability is based on pilot testing of the measure. One hundred and one charts were pulled from three practice sites. There were two observers who came to the majority of agreement, 96 percent based on the denominator and 75 percent based on the numerator for the reliability.

That was where the concern of the workgroup came in. So the workgroup expressed concerns, we didn't know how those 101 charts were pulled. Were they randomly sampled or how were they selected?

Two, the workgroup had a concern about the measure requiring a minimum of two encounters within the measurement period before a patient is included in the

denominator. And they questioned whether one visit is more appropriate in the management of individuals with MDD.

The last concern was about variability in the way in which providers assessed this and that potentially use of a standardized tool would help that situation more. Bonnie, you may have other comments.

MEMBER ZIMA: Yes, I think just to add anything additional, on the specifications what's interesting is the way suicide risk is operationalized. And it's a little bit different than how suicide risk is operationalized in the NCQA measures.

And in this measure it includes identification of specific psychiatric symptoms such as psychosis, mania, substance abuse and, quote, medical conditions that may increase the likelihood of acting on suicidal ideation. And so the implications are twofold. One is that typically in a standard psychiatric evaluation, you do rule in and

rule out things like psychosis, depression, substance abuse and medical conditions. So the implication of having that clause is that you might overestimate suicide risk in this measure.

The other issue almost bordering on feasibility but I'll mention it here because it's related to specifications. And that is that, how do you operationally define, quote, medical conditions that are not -- that are going to -- may increase the likelihood of acting on suicidal ideations? So I think that in thinking about -- you know, in actually using this, how would I quantify that like in a chart, record, abstraction tool?

The other issue that was not mentioned yet was again with feasibility. In this -- in the materials presented, there's an assumption that all data elements are in the EHR, and that's stated under their response on 3(b)1. But what's interesting is when you look at their pilot data of the 101 records,

there's actually 75 percent where the abstractors only agreed on the numerators. So 25 percent of the time they couldn't agree on the numerator. And so I think that also kind of raises a little bit of awareness of maybe some of the difficulties in operationalizing what this suicide risk is.

And then the third point is, and I think it's important, is that they do disclose that there's no data on performance, despite use in federally funded programs.

CO-CHAIR PINCUS: Any comments from the developers?

MS. TIERNEY: Yes. So you've raised a number of issues so maybe if we could just deal with them individually one by one.

So first, I think Dr. Briss, you mentioned it earlier and you just mentioned it now about the numerator and how it's defined. So the numerator is defined loosely and I think some of the items that you are concerned about, particularly the identification of

specific symptoms, we actually are -- allow -- we're not prescriptive with how the risk assessment needs to be completed. And we've provided some guidance based on the guidelines and what the guidelines recommend for a suicide risk assessment but that's not a requirement.

We also acknowledge that there is -- there are standardized tools to do this and we specifically mention the Columbia Suicidal Severity Rating Scale however it's also not required. It's just one option. And so the -- some of the information you were quoting is guidance but not required by the measure.

And the primary reason that we followed that model is because, you know, in talking about this measure with our workgroup they felt that we didn't want to be overly prescriptive. I think someone, and I'm not sure who, when I was sitting in the peanut gallery, I overheard someone over here mention sort of a concern about, you know, quote-

unquote cookbook medicine or overly describing the way such assessments need to take place. And so we felt that it was appropriate to leave it up to the individual practitioner and based on the needs of the individual patient, the type of assessment that's conducted. So that I think addresses -- at least that's our perspective on the numerator statement.

I know you had some other questions about the feasibility and the ability to capture the data in an electronic health record which I think Kendra could probably speak to better.

MS. HANLEY: I think that our pilot testing experience shows that it can be captured in a variety of different ways. And what we provide, especially because this is a measure that has been implemented in national programs, is a standardized way in which to report the data. You know, we can't actually provide a specification that could be tailored to every individual EHR out there so hopefully

that will address the issue about the presence of the data in the EHRs. It's really up to each individual site to tailor their workflow and then use the standardized terminologies to report the data.

And then to the issue about the requirement of two encounters for the denominator, you know, that's something that is a standardized approach that we've taken at the PCPI when we're looking at a measure for a chronic condition. It establishes a level of accountability for that individual physician. It's not to say that you should only assess suicide if you know you're going to see the patient more than twice in the year, you should really be doing that at every visit. But for purposes of actually calculating the measure and reporting on the performance rate, we do require that minimum of two visits. So that's kind of the background and rationale around that approach.

CO-CHAIR PINCUS: So just to

clarify, so where do I find the actual specifications that are used for determining the numerator?

MS. HANLEY: So we attached the eMeasure zip file with the submission form. I would ask the NQF staff if they could help point the committee members.

CO-CHAIR PINCUS: Is this only an eMeasure or is it also --

MS. HANLEY: It's only an eMeasure.

MS. BAL: Yes, so that would be in SharePoint. Let me just --

CO-CHAIR PINCUS: My thinking is that would be useful for us to just see what the specifications are.

MEMBER ZIMA: You know, because I think when we talked about it in the workgroup, you know, we totally got it that, you know, that you were adapting something that was already used for other chronic medical conditions. But I think what came up in the discussion in the workgroup was that

what varied on this one compared to let's just say monitoring hypertension is the higher risk of lethality. And that's sort of what got the workgroup a little bit more concerned.

MS. HANLEY: You know, I think this is something that in its current state, this is how the measure is implemented. It's something that we can certainly take back and consider for future updates. It's a minor update that could be made so we certainly hear the concern about that requirement.

CO-CHAIR PINCUS: While we're waiting for them, somebody to bring up the specifications on this -- on the screen, why don't we go around and ask for comments. I see Mady and I see Bob. Other people who have comments?

Mady?

MEMBER CHALK: I just have a couple of questions.

One, is there more than one standardized tool for suicide risk assessment?

DR. PIERCE: There's one that's been more studied than others and that's the Columbia Scale. But there are others out there that other health plans use.

MEMBER CHALK: The reason I ask the question is that it's my recollection that in many of the screening measures that have come through this committee before, the big issue has become, oh, you should be using a standardized tool of some sort. I mean, we have four or five when we talk about screening for substance use disorders. They are standardized.

So this measure we're saying, oh, it doesn't matter and that concerns me. That's number one, apparently.

The other question I have is about EHR specification. Was this tested in three EHRs as is NQF's current statement about other measures?

MS. HANLEY: So thank you for that question.

It was tested in three EHRs and I would ask my colleague Meredith Jones who is on the phone to maybe provide a little more detail about the three different testing sites.

MS. JONES: Hi, good afternoon -- or good morning, everyone. Meredith Jones. Thanks, Kendra.

This measure was tested at three different EHRs, at three different sites representing different geographic locations across the country. They used three different EHRs. All of the sites were able to successfully implement the measure's specifications that were provided to them. Some sites did use the PHQ-9 as their validated tool and other sites used -- the two other ones used the Columbia Suicidal Severity Risk Rating Scale.

When the measures were implemented at each site, they did have a conversation after changes were made to the EHR, the

programmer and the physician champions had conversations with the clinicians who were going to be documenting moving forward using this measure in the EHR kind of talking about the workflow and what they would be doing. And once all the clinicians got together, they all successfully were able to report on the measure and have since been using the measure since 2012 in their EHR.

MS. HANLEY: Thanks, Meredith.

MS. JONES: Thank you.

MS. HANLEY: And Poonam is -- if I can just explain the EHR specification that is being displayed, this is the standard HQMF eMeasure format which is an HL-7 standard for representing a quality measure specification. This is the format that all measures included in the meaningful use program are presented in and it adheres to all the national standards that have been adopted for representing quality measures for electronic data sources.

There's a top part which explains

some background about the measure; it's called the header. And then if you scroll down --

CO-CHAIR PINCUS: Where is that on here?

MS. BAL: It's on the SharePoint site. If you go to --

CO-CHAIR PINCUS: Right. But are you able to bring it up so that we can all see it?

MS. BAL: It's brought up. It's on -- that's the document in front of you.

CO-CHAIR PINCUS: I'm looking at it but I don't see anything that has specifications.

MS. HANLEY: So the top part is the measure header which includes who developed it, the measure description --

CO-CHAIR PINCUS: Right.

MS. HANLEY: -- the language, the exclusions, the exceptions. If you scroll down to the population criteria, this is where each of the data elements are listed.

CO-CHAIR PINCUS: Right. And so can we see the data elements for the numerators?

MS. HANLEY: Sure. So it's identified as an intervention performed suicide risk assessment. That data element maps to a SNOMED concept for suicide risk assessment which again is consistent with our more open, less prescriptive approach to what we require for reporting.

CO-CHAIR PINCUS: So what terms are included under that?

MS. HANLEY: Let me get to that.

All of the value sets that correspond to the data elements for the eMeasures are hosted publicly by the National Library of Medicine on what's called the Value Set Authority Center. And --

CO-CHAIR PINCUS: Yes. But I guess my concern is that, you know, we're sort of like buying a pig in a poke. You know, we need to be able to see what those terms are so we know what's being specified.

MS. HANLEY: So it is one -- it's one SNOMED concept from the procedure hierarchy, and the concept description is suicide risk assessment.

CO-CHAIR PINCUS: And so I'm not an informatician.

MS. HANLEY: Sure.

CO-CHAIR PINCUS: So can somebody explain to me in layman's language exactly how this data is captured and what terms constitute a suicide risk assessment?

MS. HANLEY: So it would really rely on the individual physician stating that that's what was done. So again, that's --

CO-CHAIR PINCUS: So they have to use the term suicide risk assessment?

MS. HANLEY: The information captured in the EHR would be mapped and then reported.

CO-CHAIR PINCUS: Does it have to be a separate field?

MS. HANLEY: It could --

CO-CHAIR PINCUS: Do you use natural language processing or could it be a separate --

MS. HANLEY: It could be a separate field. It could use natural language processing.

CO-CHAIR PINCUS: So what terms would count?

MS. HANLEY: Any of those would count.

CO-CHAIR PINCUS: Any -- what do you mean, any of which? So if I wrote assessment, it would count?

MS. HANLEY: So that's where we use the guidance in the definition. So it would be in accordance with the guidance presented in the definition --

CO-CHAIR PINCUS: So how --

MS. HANLEY: -- if you meet -- and this goes back to what Sam was speaking to earlier about it, to be at the discretion of the individual clinician and specific to the

needs of the patient.

CO-CHAIR PINCUS: I'm still -- okay, Bob?

MEMBER ATKINS: I have the exact same concern and I'm old and don't know everything you're talking about. So let me take it back to paper records. SI equals, and have a zero and a line through it. Does that count as a suicide assessment? Suicide risk assessment? Suicidal ideation, SI, equals zero with a line through it. Would that count as the person did a good job?

CO-CHAIR PINCUS: No.

MEMBER ATKINS: That's, I think, unfortunate.

(Laughter.)

MEMBER ATKINS: Because that's not suicide risk assessment, that's a conclusion because they have on their form that they have to address whether or not the person had SI. And if suicide risk assessment is tell me about all the things that you have in your

verbal description where a person actually acted like a clinician, and I don't -- I'm sorry, I'm not trying to be rude, but I just -- I've looked at a lot of paper records and I would not be comfortable with what I just -- when I heard that.

And I'll stop talking. People are starting to laugh at me, so --

CO-CHAIR PINCUS: No. No. I think --

MS. HANLEY: That's not what the definition is.

MEMBER CHALK: Yes, that's what I meant when I said somebody could do anything. That there isn't a requirement of the standardized tool which --

MEMBER ATKINS: I'm sorry, even without a standard -- I mean, I also know a lot of people aren't going to be doing a clinical interview that will address the issues that -- I'm just like a bunch of people talked about how do you assess suicide risk.

Any one of them is good. But -- so sort of three ways to do a good job. There's a standardized tool, there's a good clinical assessment or there's the SI equals zero. And I think the third one shouldn't count, but the first two you're saying would count. But you also -- the third one would count and that's my concern.

MS. TIERNEY: I think one of the challenges is that, you know, these are performance measures so we have to be able to use data and information that can be assessed in a uniform way. And so you know, specifically for an eMeasure, you know, some of the things -- how do you define a good clinical assessment? And if -- even if you do define it, let's say we defined it as our definition does, which provides some information and guidance to an implementer, are all of those things going to be things that we could code and identify data elements from an electronic health record?

So I mean, I think some of this speaks to just the challenges in performance measurement in general and the ability -- the need to develop measures for which data can be collected and captured and analyzed.

I will say just kind of more generally to your point earlier about for other screenings, we have very clearly defined systematic tools. I would say that for -- you know, we at the PCPI have developed other measures around alcohol use screening and we do refer to specific tools because I think those have been very well studied and there's a very clear evidence of their effectiveness and their utility in identifying and screening patients who are at risk.

I don't think that the studies have been as well conducted or as well documented for suicide screening and the USPSTF sort of references that in their recent article. And so I think it would be very difficult for us to prescribe one particular tool that may also

be potentially burdensome and time-intensive. I will say for tobacco screening we do not have -- we also have developed the measure that's in the meaningful use program and in widespread use and we haven't defined a particular type of tobacco screening that has to be conducted because of, you know, wanting it to be something that could be of relatively low burden. And I think if we were going to prescribe, there probably are tools for tobacco screening assessment but we have wanted to use something that could be feasible to capture and reasonable to implement.

CO-CHAIR PINCUS: Let me -- what I'm trying to get at is I just want clarity. I'm not taking one side or another. And so -- and to know exactly what counts and what doesn't count. If I want to implement this in my setting, is there a methodology that allows me to do that? So you know, if this definition up there is a long list of different items that you mentioned: inquiries, specific

inquiry about suicidal thoughts, attempt plans, means and behaviors, identification of specific psychiatric symptoms or general medical that may increase the likelihood of acting on a suicidal idea. Assessment of past and particular recent suicidal behavior, delineation of stress and potential protective factors.

If I did any one of those, that would count?

MS. TIERNEY: Yes.

CO-CHAIR PINCUS: So if I asked, do you have a history -- if I asked about psychosis, do you have a history of delusions or hallucinations, that would count?

MS. TIERNEY: So we are not prescriptive. So some of this does rely on --

CO-CHAIR PINCUS: Unrelated to suicide, I just asked do you -- you know, have you ever heard things that --

MS. TIERNEY: There is some sort of --

CO-CHAIR PINCUS: I mean, no, is it yes or no, would that count?

MS. TIERNEY: So I guess you could say this with any measures, right? Because --

(Laughter.)

MS. TIERNEY: -- there is -- yes, there is - I mean, I think that with performance measures in general the issue of gaming comes up. You know, oh, I could just check this box and I haven't actually done what I said I was going to do by what the measure prescribes, but I check the box and so it counts.

So I think most of the studies that I'm familiar with that have kind of assessed gaming in general with performance measures have --

CO-CHAIR PINCUS: I actually wasn't asking about gaming. I was trying to see if -- I wanted to do this, but I have to sort of program, you know, ask the people sort of in my informatics group to program it to count

somebody asking about hallucinations or delusions independent of any issue around suicide.

MS. HANLEY: I mean, the general medical conditions that are described are examples of things that may increase the likelihood. So I would say asking about hallucinations, that would not count because that's not asking about suicide and what --

CO-CHAIR PINCUS: But I guess you would say that, but is there any way that I -- I mean, how would I --

MEMBER LARDIERI: Could I jump in for a minute?

CO-CHAIR PINCUS: Yeah.

MEMBER LARDIERI: Because I live in the EHR world so in order to do this, Harold, I don't think you're going to be searching the EHR for each one of these things. What you're going to do, the provider is going to do what they do, ask one or all of these things and then there's going to be a check box that is

going to say suicide risk assessment. They're going to check that, then that gets correlated to the SNOMED code that then says yes, he is.

CO-CHAIR PINCUS: So what you're saying, that there has to be a field?

MEMBER LARDIERI: There will be a field, the checkbox in the EHR.

CO-CHAIR PINCUS: Again, I'm not -- I'm just trying to -- my understanding is that it doesn't have to be a field.

MEMBER LARDIERI: There's no way you could do without having a field unless you can do a language processing against this stuff. And you're not -- we're not there yet, unless you use a Watson.

CO-CHAIR BRISS: I think we may have established -- at this point, it seems to me that we've established that there's some potential looseness in the definition. I think we ought to move on, right?

CO-CHAIR PINCUS: Yes. And I guess the other point here is, are we really doing

a risk assessment or are we just asking about ideation? And I think that the state of the art of the evidence and what we have to grade risk and accurately predict, particularly given the relatively low incidence, not trivial obviously. But I mean, to really do the study and to be able to look at varied ways of assessing risk and then looking at attempted and completed suicide is a study that I don't think has been done in a broad approach to assess what are the relative ways one can assess suicidality and its predictive ability over time.

Let's -- people have their hands up or their cards up, so Julie, Tami, Mike and David.

MEMBER GOLDSTEIN GRUMET: So I think one of the things that I'm -- I was struggling with is what you said, is this about doing a screening tool or doing a risk assessment? Because there needs to be a screening tool in place which is not necessarily the same as

assessment. Is the person with major depressive disorder having thoughts of suicide, period, and there are many tools available that function as screeners. And that's -- that piece seems to be missing.

I agree that the whole argument about this, you know, list, which is an assessment, is the next step after a person is deemed at risk. And there are standardized screeners and you could provide a list of standardized screeners.

My other concern a little bit is that we're limiting it to major depression. And in kids, often it's anxiety and it's psychosis, substance abuse. So I wonder why we're not including them as well. I want this tool, I really -- you know, I really believe it's very important that kids are assessed for suicide risk when they're struggling with mental health issues. But I think it should be broader and I think providers need a lot more guidance because I think they're not

going to know how to ask these questions and we're going to struggle with a lot of providers saying you're not thinking of killing yourself, right? And we know how inappropriate that is.

(Inaudible comments.)

MEMBER GOLDSTEIN GRUMET: Well, but I think that's what we know, is that providers do and they don't know -- they're not trained well which -- you know. But I think we need to give a little bit more guidance and also clarify, is it screening or is it risk assessment? And both should happen.

CO-CHAIR PINCUS: Tami?

MEMBER MARK: Moving a little from the discussion of how prescriptive the numerator is or should be to how statistical the denominator is. Do you have specifications for MDD and are those required as part of this use of this measure? Are you able to put those up, too?

MS. HANLEY: So the denominator is

specified through diagnosis of major depressive disorder and the value sets for those are specified in ICD-9, ICD-10 and SNOMED. I don't know if Poonam has access to the VSAC. I have them on my screen but --

CO-CHAIR PINCUS: Okay. While they're bringing it up, Mike?

MEMBER TRANGLE: You know, I have some questions just about inter-rater reliability which kind of merge with feasibility. So the concerns sort of center around there are 102 patients that were part of this background thing in three different locations, you know, one being a physician-owned practice, two sites, one being a community mental health center and third one was sort of a more primary care center. And I'm quite concerned as to if you extrapolate this to the real world, and maybe this was real world versus kind of places than we're used to doing research and weren't like family practice docs, 20-minute visits putting

everything in, you know, I'm interested in how the uptake in the utilization and reliability vary between the three different sites. Do you see what I'm asking?

I have an intuition saying a busy community mental health center with practitioners that aren't used to doing research are not necessarily going to get into this, you know? Same thing with family practice. And I'm just kind of wondering about, you know, vary from sites -- are there hints in the mental health world that may play out better than in primary care or vice versa?

MS. TIERNEY: I'm going to ask Meredith. I don't know if you can speak to the testing project at all, and specific to the question. I don't know if that was assessed.

MS. JONES: Yes. So it's a part of the PCPI testing methodology to report one kappa score among the sites to demonstrate reliability. If you're interested in looking

at the final results from each site based on the number of charts pulled from each site we can certainly provide you that information.

I will reiterate what I said earlier and it kind of gets back to the conversation we've been having, is that suicide risk assessment was implemented into the EHR in a structured field. So it was, you know, like a checkbox, it wasn't something that was living in an open form or unstructured field. The sites each were able to implement the data elements in the measure that you're seeing on the specification in structured fields. But back to the point, if you would like to see the final results and the differences amongst the sites, we can provide that information to you.

MS. HANLEY: So it sounds like the question was more looking at how does this actually get implemented in real life, in an everyday clinical scenario. These were sites that were willing to help us test the measure,

to test the scientific properties of the measure. I think we'll see as broad implementation and uptake of this measure progresses, you know, at that point we'll be able to have more feedback at how it actually -- how sites have adapted to incorporating this measure.

MEMBER ZIMA: I just wanted to follow up on that question. I don't think they had the statistical power to test across the three sites, and only 15 records were contributed from the community mental health center.

And the other thing, too, which continues to be a question from the workgroup is the sampling of those three sites. Because one site was a very large extensive network of health centers; one was one private practice with two locations, a suburban location and urban, so I'm assuming the provider was one, right? And then we had CMHC. So I -- that's what I wanted to say.

MS. JONES: There were two practice family -- two physicians from site B, the second site.

MEMBER ZIMA: Okay. Fair enough.

MEMBER EINZIG: So giving clinical perspective here, what we do in our clinic is we ask the question, are you suicidal. And before they see us in clinic they fill out a sheet and on the sheet it asks about suicidal thoughts or any safety concerns.

So thinking simplistically, big picture, clinical perspective, what we're trying to do is we're trying to prevent suicide. That's what this measure is about. And raising awareness, asking the question.

There's plenty of screening tools for depression and then, you know, I remember reading an old study that if you ask the simple question are you depressed, they're going to answer yes or no. That just has the effect -- just as valid as any depression screening tool out there. So if you ask the

question, you know, I think that's reasonable to say that that's -- what we're trying to do is we're trying to make -- measure the quality of clinics in asking about suicide and preventing suicide. So if it's as straightforward as documenting that with no suicidal ideation reported or no safety concerns, I think that's reasonable and valid.

CO-CHAIR BRISS: Yes, that's what I -- it's going to give us more primary care perspective. I think that we could move the ball forward significantly by encouraging people to ask that question. I'm reasonably comfortable with a simple checkbox in this context, given where the field likely is, and I just wanted to validate what Bonnie said. I that this is an important enough issue that the usual chronic disease rules at PCPI shouldn't apply and I think one -- this ought to be enough.

CO-CHAIR PINCUS: Actually I'd like to sort of step out of the Chair role and just

make a concern in that we're doing this as an accountability measure. I mean, because that's -- you know. So we are placing on the same level of accountability -- and I'm thinking about this because it -- in our hospital we're actually implementing the Columbia Suicide screening tool. And we are sort of encountering some issues in implementing it because it's -- if you look on our inpatient settings there is an extensive suicide assessment that's already in place. And this is sort of on top of that.

And so on the one hand, you know, we have sort of that issue of people feeling sort of an additional burden, on -- but also, you know, we are really, you know, taking this very seriously and we're really running a full-scale suicide assessment either -- you know, either a very complete one as we have sort of traditionally done, or the Columbia one. Yet we are going to be compared -- which comprised a considerable amount of effort in

implementation. And we're going to be compared with other practices that all they have to do is have, you know, SI equals zero.

And it seems to me that there is an imbalance there as an accountable -- accountability measure. And so that it's not just an issue of consciousness raising, but it's -- you know, if there's going to be skin in the game on these kind of things, it seems to me there should be a standardized level of expectation.

People raise their hands who -- just do people want to -- Jeff?

MEMBER SUSMAN: I just was going to say that, at least in my experience in looking at what people in primary care are actually doing in our research, actually observing encounters with depressed patients, this is a really woeful state of current practice. My belief is that in most primary care settings where this isn't a particular focus, that suicidality is not assessed in any form most

of the time, in fact, the vast majority of the time, which I think is a real problem.

There are lots of issues with this measure, I get that, and I agree, Harold, you could be unfortunately compared to the person who just checks the box. But I think the greater good here is that we're starting to measure this and to look at it and that it will be on health system screens. So I'd say let's not let the good enough, you know, stand in the way of making some progress here, recognizing that there are clearly some issues with this.

MEMBER MARK: Does anyone know the frequency of MDD in primary care as opposed to the question NOS? I mean, I'm concerned that maybe the sample size that we have in these facilities is so small as to not be meaningful. And I guess, you know, in terms of your point about, you know, weighing, comparing, specialty facilities to primary care facilities, you know, again if they're

such a little sample, you know, how useful is this measure going to be? As opposed to maybe for specialty centers, okay, you have enough to be useful. But in primary care settings where it sounds like we're really trying to make a difference, the denominator may be too small to be useful, to make this a useful measure.

MS. JONES: Hi, this is Meredith Jones again.

I just want to share a little bit more about the PCPI methodology. I'm not sure since our last conversation if you've been able to read a couple months old document which we sent to NQF staff about the testing/sampling methodology we use. We use the Donner Eliasziw kappa sample by population to determine appropriate baseline number of charts to abstract for each measure.

This approach we use uses the two-tailed test to determine significant sample sizes, we practice this in each measure

testing project. We use a value of the expected proportion of positive ratings for a measure to be tested based on available data on average performance clinicians at each site on the measure. So for example, if the average performance would be ninety percent, the proportion of positive ratings is point-nine-oh, and we use that two-tailed test at 80 percent power to detect the difference between the value of a calculated kappa, which you see, and the null value of a kappa.

And again, I'm not sure if you got our additional guidance and the methodology that we used but the final sample of 101 is statistically significant.

MS. HANLEY: This is Kendra.

I also just wanted to comment on the point about the relevance to this in primary care if the sample is so small.

We're also operating in an environment of accountability where all practitioners are in need of measures to

report to participate in many of these public reporting programs that are affecting their reimbursements. So it's also a measure that's very important to those mental health providers who are treating patients and do have the proper sample size.

CO-CHAIR PINCUS: So I think we're ready to -- Bob?

MEMBER ATKINS: I'm sorry. I guess one final thought.

To me, a lot of the issues here would be reconciled if we didn't set the bar at assessment. A lot -- some of the comments seem to be around screening. The one question I think have to ask is, SI, yes or no? That in my mind is not an assessment, it's a screening question. I think by labeling it assessment you're setting a different bar. And so if you drop the bar to screening, then doing the full clinical and the Columbia and everything else is a very rich screening tool, if you will.

But the one question would also count as a screener and maybe that would help. I mean, it would help me at least. I had a whole different bar in my mind until this conversation.

MS. HANLEY: So again, I think we'll take that feedback back and we can consider that for future updates.

MEMBER TRANGLE: This is a general comment that I think we should think about, not just for this measure but all measures. You know, this is a new standing committee and the point of having a standing committee is to sort of vote, say yea or nay, but also to take part in process improvement of the measures over time. And I think for everything we're talking about we may need to have a standing agenda item for, you know, we vote yea or nay but then we have recommendations for measure improvement that they'll come back to us with.

And then if somebody could actually pay attention to half or a third of our

comments for the measurement improvement, you know, and then summarize that for us so that somehow we could vote at the end about these are the key ones you should work on and come back to us, it would make us more efficient.

CO-CHAIR PINCUS: So let's move on to voting.

MS. BAL: Okay. Voting for reliability for 1365 is now open. And the options are one low, two moderate, three low, four insufficient. Sorry, I said one -- I meant one high. I was looking, I was like, I feel like I said low twice.

(Laughter.)

MS. BAL: Okay. The results for reliability for 1365 is high three, moderate twelve, low three, insufficient six. And yes, we're good to go forward with the next. And then we're -- yes, above 60 percent. So and now the voting for validity is open. Oh, that's odd, sorry. Okay, now it's open. And the options are one low -- I'm sorry, one high

-- one high, two moderate, three low, four insufficient. One high.

MS. BAL: Okay. The results are high one, moderate thirteen, low four, insufficient six. And we will move forward with this measure to feasibility.

It's 24 votes instead of 25 this round, and we had fourteen listed as the -- give us one second.

Forty to sixty percent is considered consensus not reached. So we will move forward but -- we continue to move forward but --

(Inaudible comments.)

DR. BURSTIN: Yes, this is the validity.

MS. BAL: Yes.

DR. BURSTIN: We'll come back to show you the percentages on validity. So, right, so that's 58 percent in our current rules, we say that if you're in the gray zone between 40 and 60, we'll continue your

evaluation and maybe additional information provided by the developers you can consider afterward.

Sixty is to move forward, right. So this is still in the gray zone and what we do now as part of our gray zone analysis continue to move these measures forward and let you finish the analysis so you don't have to then figure out how to go backwards if you get additional information.

MS. BAL: So the only way we'll fail a measure is if it's less than 40 percent. If we're in between 40 and 60, it's gray zone and we'll just document that consensus was not reached on that portion of the measure and we'll move on to the next option.

DR. BURSTIN: We can examine the public comments that come in when we release the report.

MEMBER SUSMAN: Is there any pragmatic importance to us, whether it's moderate or high consistency? In other words,

okay, so this one had 58 percent, the other one had 66 percent. It sounds like we're just doing the same thing and ultimately we're going to vote overall and it doesn't matter. Or maybe I'm missing something?

CO-CHAIR BRISS: What's the practical difference between consensus not reached and consensus?

DR. BURSTIN: Well I mean, I think the major difference is you want to identify for the public and the membership in particular that, when they see this report and it comes out, where you, in fact, couldn't reach consensus on where there were issues. So that will be clearly labeled as consensus not reached.

CO-CHAIR PINCUS: So it's a matter of public --

DR. BURSTIN: Yes.

CO-CHAIR BRISS: Public transparency as opposed to --

CO-CHAIR PINCUS: -- communication?

DR. BURSTIN: And some of it is, in the work we've been doing with our board on defining consensus there was a sense that you're just kind of creeping over the 50 percent line probably wasn't enough. So for now we've set the threshold at 60, this is obviously a squeaker. My suspicion is this will probably be moved forward. But we just want to make very clear to the people about the discussions you had.

MEMBER TRANGLE: In the end it's still listed as endorsed or not endorsed at this time?

DR. BURSTIN: Well you know, you're still really early in this process so it will go out for comment and you'll have an opportunity to re-engage in it and see if you want to reconsider any of these issues. So at this point we just -- we'll allow the rest of the evaluation to move forward but it will clearly go out with a note that this particular element on validity was consensus

not reached.

CO-CHAIR PINCUS: Let's move on with feasibility. So Bernadette and Bonnie?

MEMBER MELNYK: So not to beat a dead horse, but the concern regarding this was, again, the variability and how people assess it, and the documentation.

And then the second comment from the workgroup was just that the developers consider expanding the measure in the future to include persistent depression in the DSM-5 as well as other comorbid conditions.

CO-CHAIR PINCUS: Bonnie?

MEMBER ZIMA: No additional unique comments.

CO-CHAIR PINCUS: Other comments from around the table in terms of feasibility?

(No response.)

CO-CHAIR PINCUS: Okay. So I guess we're ready to vote.

MS. BAL: Okay. Voting for feasibility for 1365 is now open.

(Brief pause.)

MS. BAL: And the options are one high, two moderate, three low, four insufficient.

(Brief pause.)

MS. BAL: Okay. The results are high two, moderate thirteen, low five, insufficient four. And with fifteen we will move forward.

CO-CHAIR PINCUS: Okay. Now comments on usability and use.

MEMBER MELNYK: Our workgroup noted that the measure is recently in use in several reporting programs and performance data is not yet available.

MEMBER ZIMA: No additional unique comments.

(Brief pause.)

CO-CHAIR PINCUS: Okay. I guess we're ready to vote.

Oh, Bob, do you have a comment?

MEMBER ATKINS: No.

MS. BAL: Oh, sorry. I'm just going to restart it.

Okay, are we ready to vote? Voting is now open for usability and use, for 1365. Options are one high, two moderate, three low, four insufficient.

(Brief pause.)

MS. BAL: And we are -- is everybody voting? We should -- we still have one more in the room that we don't have.

(Brief pause.)

MS. BAL: Yeah, we're missing one in the room. We're at 23 and we should be at 24.

We're good in the room. Thank you.

Okay. The results for usability and use for 1365 are high four, moderate ten, low five, insufficient for five. So we're in the gray zone for this measure as well but we will move forward to the final vote.

Would you like more discussion or just to vote?

CO-CHAIR PINCUS: Any further

discussion?

(No response.)

CO-CHAIR PINCUS: Any new or unique comments?

(Laughter.)

(No response.)

MS. BAL: Okay. In that case, overall suitability for endorsement is now open for voting. And the options are one, yes; two, no.

(Brief pause.)

MS. BAL: Okay. The final result for 1365 is yes fifteen, no nine. And this measure is recommended for endorsement at this point in time.

CO-CHAIR PINCUS: So are we ready to break for lunch?

MEMBER MARK: Yes, we are. I just want --

MS. DORIAN: No, we're not. We are actually two measures behind. Perhaps we could do the next measure, at least begin it.

Lunch was scheduled for 1:10 and I think that's when it will be ready. So we can at least get started on the next measure.

MS. BAL: I don't think it's ready, though. Oh, it's back there already? Oh, they said it would take a while. We're just going to check.

CO-CHAIR PINCUS: Yeah. So why don't we make it so that people bring their lunch back here and let's take -- yeah, let's keep working while we eat. So let's take ten minutes to get your lunch and bring it back.

(Whereupon, the above-entitled matter went off the record at 12:45 p.m. and resumed at 1:12 p.m.)

CO-CHAIR BRISS: So I have us at about 45 minutes or so behind and with eight measures yet to go this afternoon. So I'd sort of like us to get restarted, please. So if everybody would be seated and get ready to go, I would appreciate it.

So Sarah, let's do the teeing up of

the set and then I promise I'll give you some pauses in which to eat and reduce Helen's maternal instinct, okay?

MS. HUDSON SCHOLLE: Thank you, Peter.

CO-CHAIR BRISS: So I lied, we were remiss at the end of the morning of not asking for public and member comments so we'll do that now, please.

Operator, can you open up the lines for public or member comment, please?

(Operator speaking.)

OPERATOR: There are no comments at this time.

CO-CHAIR BRISS: Thank you.

And with that, NCQA. Health Screening and Assessment for People with SMI

MS. HUDSON SCHOLLE: Hello everyone, I'm still Sarah Scholle.

(Laughter.)

VOICE: We are still the committee

so there's a symmetry.

(Laughter.)

MS. HUDSON SCHOLLE: That's nice.

And so actually I'll be with you for the rest of the afternoon, I think, on -- and a good part of tomorrow as well, to talk about the measures that we've developed with Mathematica Policy Research. This is under contract from the Assistant Secretary for Planning and Evaluation and the Substance Abuse and Mental Health Services Administration.

We actually began on this journey, I was going to say three years ago but actually more like five years ago because we started with a contract from ASPE to develop measures for people with schizophrenia and we brought those measures to this committee, I think about three years ago. Those were measures that focused on care for people with schizophrenia and whatnot that looked at both continuity band of psychotics and then also

care for chronic health conditions and we -- in looking at -- trying to look at the healthcare for this high risk population.

And the committee at the time said this is a really important area and we want to encourage you to continue to work in this and to broaden the work to focus on a broader set of people, not just people with schizophrenia but others with serious mental illness. And also to consider outcome measures at the time that we were limited to claims-based measures. And so that's what we did.

Now in addition to the input from this committee, we also had -- we started this work by conducting eight different stakeholder focus groups to get ideas about what was important, where we should focus our attention on developing behavioral health measures. And in it -- and we heard concerns about a variety of areas that we investigated but in particular concerns about early mortality of people with serious mental illness and lack of

access and attention to their general medical needs. So that's -- it's from both of those areas of input that we have come to you this meeting with eleven measures that look at health screening and attention to chronic disease for people with serious mental illness and people with alcohol and drug dependence.

Our theory about how to approach this set of measures was to focus on areas where we knew -- where we had evidence that there was a higher prevalence of the risk factor of a condition or risk, so a condition like diabetes or a risk like obesity in the population. And also evidence that there was a disparity in access to evidence-based care. And that's how we -- we actually worked through the literature, we looked at guidelines both for the general population and guidelines for the populations of people with serious mental illness. We looked at evidence about risk and disparity and of course the risk sometimes is based on the treatment and

-- as well as the condition.

And we wanted to focus on existing measures and to think about how we could use the existing measurement enterprise to shine a light on a high risk population. So that these measures are actually -- all the measures we'll talk about in this group are focused -- are measures where we started from an existing measure for the general population, we looked to see was there evidence of a high risk that people with SMI or AOD were at high risk for the condition or had a disparity in care. And then we looked to see whether the measure numerator needed to be altered. Would you expect some kind of difference in care for this high risk population?

We conducted testing in health plans, the -- you know, we considered who would be the right -- where would be the right level of accountability for these measures and we focused on health plans and -- thinking

about health plans and states are higher levels as being responsible because an individual behavioral health provider or an individual primary care provider might not have all the information they need or all the access to be able to be responsible for making sure that somebody with a serious mental illness gets BMI screening and follow-up. But health plans should be able to be responsible for that.

So we tested the measures in three health plans. All the measures are specified for using administrative data, claims data to identify the denominator, and a chart review either of an electronic or paper chart to determine the numerator. And our results showed in particular for the measures we're going to consider in the next series, the obesity, diabetes and hypertension measures, we saw big disparities for this population compared to the general population. And we had extensive public comment and we conducted

focus groups with stakeholders to get their reactions to our information and ideas about feasibility and usability.

Now we have submitted these measures as individual measures as they were, so we've specified them as individual measures even though we present them as a group, we present them as a group of measures for people with SMI. And part of the reason for presenting them as individual measures rather than as a composite where we'd say that people with SMI get everything or -- is that we wanted to allow for flexibility of implementation. So there's two ways that these measures could be implemented to try to improve the usability by different organizations. So a health plan that is looking at diabetes care for their general population could over-sample and report the diabetes measures for people with diabetes and serious mental illness, okay? So they could lessen the cost of doing these hybrid measures, doing these chart reviews by

doing that over-sampling. And then they'd really be able to compare what does it look like for the general population? What does it look like for people with SMI?

Another alternative would be to say, you know, what we really want to do is say look at people with serious mental illness and look at all of their risks and their needs for screenings and for attention to chronic disease. So you could say let's start with people with SMI and then out of that group we're going to expect a good proportion of them to have diabetes or hypertension, so we could do the diabetes and hypertension within that group.

So that's why we presented the measures as individual measures for your review. If they stand alone as an individual measure then that allows organizations that would implement these measures, whether it's -- they're measures that could be implemented for health plans or measures that could be

used, considered for the core sets. There's more flexibility when they're an individual measure and we thought that that was important to achieving implementation to allow that flexibility.

So I think that's the introduction. Thank you.

CO-CHAIR BRISS: Thank you.

And any questions or comments from the committee?

(Brief pause.)

CO-CHAIR BRISS: David, would you start?

MEMBER PATING: Dr. Briss, I just would like to ask, it's kind of a process question, how we should do this. Because I like the idea that we would be looking at these separately. I mean obviously we've looked at some measures and we've munched them together and we get the goulash. I heard a fruit salad, you know, it's harder to digest.

But at the same time these are

obviously, if we do consider them all, there's a cumulative impact and I just was wondering how we would handle that discussion. Would you want to look at the cumulative impact kind of at the tail end of all the discussions or should we do it like right now at the beginning and get it out of the way and then divide up the different parts?

You know, I imagine if you did this right, you'd have to -- you'd almost want to set up a whole -- you know, you do your psychiatric visit and then you come on back and you do a physical screening with a series of, you know, examinations to make sure that you do this checklist correctly. But I mean, that's not the question that's being asked here.

CO-CHAIR BRISS: Why don't we at least try to tee up those issues. Now I have a sense that there are going to be -- there are likely to be common themes that run through all of these measures, and so it

doesn't feel to me to be efficient if we address common themes kind of randomly and variably in measure one and three and seven, and then it's a really -- do you want to talk a little bit more about the cumulative impact issue?

MEMBER PATING: No. I mean, I just think that this is definitely medical and you could piecemeal it out and do it multiple visits. But probably to do it right, I mean, I would imagine you'd find the stethoscope at the back of your drawer, wherever it's been for the last 20 years, pull it out, and you know, you do a visit, you know, that systematically goes through and makes sure that you've done a physical, you've done a BMI, you've ordered the blood tests, you've set up the consultations. And you know, somehow it's framed off.

I was actually talking to Ms. -- Vanita, and she was saying, well, is it even billable? So there's even kind of questions

about, you know, the pragmatics of this systemic level. But I don't have any more comments to that other than they're already obvious -- these issues of feasibility and total burden and how they impact both the set and then the individual aspects.

CO-CHAIR BRISS: So David, is that an argument for -- is one of the net effects of that that you'd like to eventually like to see a composite? Is that -- or --

MEMBER PATING: Well, that's where I think we're wanting to go. I mean, in round two of this I think we had a measure where somebody was going to get a diabetes measure and we were saying well, what we really wanted them to do was to get a health screen, but we didn't think -- I think, if I remember correctly, and I can't remember if it passed or not. But I mean, I think that these are definitely pushing towards addressing, you know, the thing that shortens -- the 25 years that shortens the chronic life of those with

severe mental illness, right? But again, if it's -- somebody has to do this and I think in the past we said what you really want is not a bunch of piecemeal stuff, you really want them to get linked with their primary care.

And so these are just these generic questions of is it -- you know, who does this, how do we do it? Is this measure pushing us as a set in that direction? And yet I also appreciate presenting them individually because each of them has aspects that we'll have to look at.

MEMBER ROBINSON BEALE: Yeah, I just want to make a comment. I know that this is an existing provider level measure and I think by adding the health plan level brings more teeth to the measure from the standpoint of view the health plan will have more access to data that will help identify those who have not -- if everything's billed correctly, have not had those gaps in care.

And I think this is important

because, I'm just going to say, some states are moving to these quicker than you think and are starting to make plans financially responsible for these measures. So it has a greater level of importance now than it did before.

CO-CHAIR BRISS: Bonnie?

MEMBER ZIMA: Just a general comment. In thinking about these, it's actually the timing of considering these quality measures only because I think if we had integrated care models, you know, it would fly with feasibility, right? But at this point we still also have problems with missing data with behavioral health records, not being able to connect, difficulties sometimes linking records in medical primary care and specialty mental health. I think that's going to be an issue.

CO-CHAIR BRISS: And Dodi, I apologize, I missed you. Sorry. Please go ahead.

MEMBER KELLEHER: I'm just curious, the question I have is I understand the importance of this because of the disparity issue, but only in tobacco screening did you include both seriously mentally ill and substance use and alcohol. And I was wondering what the rationale was for -- was there not enough evidence for the others?

MS. HUDSON SCHOLLE: Yes. We -- that's the problem. There was not evidence that -- and actually we spent a huge amount of time looking for information that would help to establish whether there was a higher risk of problems. And in fact, we tested -- there's just not evidence that people with alcohol or drug dependence are at higher risk of diabetes or hypertension. They're --

MEMBER KELLEHER: How about the disparity issue?

MS. HUDSON SCHOLLE: We were not able to find evidence that was on a broad base. I mean we -- and we searched a lot to

look through that. It was clear we found -- it was easy to find that information on SMI.

MEMBER KELLEHER: So the disparity, did you look at disparity in terms of people with -- in treatment for, say, substance use or alcohol and/or, you know, coming from, say, the substance abuse, you know, systems, disparity in their getting, you know, what we're describing as basic health screening and treatment?

MS. HUDSON SCHOLLE: So our data clearly show that people with serious mental illness and alcohol and drug dependence don't have good access to general primary care. And we were encouraged to consider that as a measure, did people get access to primary care? But then you have to be able to say, well, what do you expect that visit to be? Otherwise it's not really an accountability measure, it's really an access to care measure and would not survive your evidence rules for NQF about what's the importance. And there's

not a guideline that says people should have a primary care visit, as sound and obvious a statement as that would be.

So that's why we -- you know, but that's why we went down the path of trying to say, okay, well, what should it be and where would we have enough evidence to say there is a risk, this needs to be addressed? So when we're looking for a quality measure that we can defend on all those criteria for importance, we felt like we needed to really try to adhere to -- to try to provide the evidence to support at each stage.

And that's why we ended up looking at the existing measures and trying to say this is -- this is actually going to be more feasible because everybody knows what those diabetes measures are and everybody knows that BMI measure -- you're already doing it. Report it for this sub-population so you can target your focus here. And what you'll find is that when you do that you'll get more

people into primary care.

MEMBER MAZON JEFFERS: So I think I have three points that I hope are overarching to your point of all these measures. The first one that came up in our workgroup was that, by proposing a set of measures specifically for the SMI population, we sort of begged the question, is there another population, sub-population or specialty population that there should be a similar set of specialized measures developed for? So that's a question that came up in our workgroup and I think it's a valid one.

Another point which -- it was helpful to hear your explanation of why you went for you know, the single measure to maximize flexibility to allow the measure to be used in different settings, either the specialty mental health -- specialty behavioral health setting or in a physical health setting. For me, I find not just for these measures but for a lot of the measures

we're talking about, I find that allowing for the flexibility actually clouds my clarity on exactly how are you using this measure and in which setting and for what purpose? You know, what are we trying to see or show or demonstrate.

So for example, are we talk -- you know, if you think about the SAMHSA four-quadrant model and the appropriate setting for people with SMI to be treated in, are we talking about using these particular measures in a, you know, quadrant four for someone who has mild to moderate -- I mean, I'm sorry, moderate to severe behavioral health issues and therefore we're talking primarily about a specialty care setting? Or are we talking about people who might be identified in a primary care setting?

So while I appreciate the flexibility, I find it also contributes to my lack of clarity because I'm not sure what setting were you trying to use the measure and

for which population and for what purpose ultimately. Is it to make the behavioral health provider take their stethoscope out of their drawer that's -- and blow the dust off of it? Or is it to get the primary care provider to recognize that there might be people with SMI walking through their front door? And I think it's a slightly different focus.

And sorry, the third thing is that particular issue of the care setting and how the measure is constructed and its relevance becomes more acute at the systems level particularly because the -- in many, many states the two systems of care are completely siloed and their data systems are completely siloed. So the feasibility of getting body mass information in a -- from a MBHO is very, very difficult at this stage.

So those are the three issues that I think overlap.

MS. HUDSON SCHOLLE: Number one, why

SMI? That's the subgroup that our stakeholders and focus group told us to look at, based on concerns about mortality.

Second --

MEMBER MAZON JEFFERS: Sorry, it's not "why SMI?" It's "why not something else?"

MS. HUDSON SCHOLLE: And we were limited in the resources that we could provide and that was the highest priority that we were directed to.

What settings? So these are measures specified and tested at the health plan level. So a health plan could implement them in a whole variety of ways. Health plans could tell primary care providers these are the people with SMI on your panel, go find them, okay? They need this screening. They could go -- the health plan could work to develop an integrated setting for care for people with SMI. Or a health plan could say we know you're the general provider, you are the place where people with SMI are, we're

going to send you a nurse there or we're going to pay you more to do this, to create this system. So that's the flexibility.

The reason we put it at the health plan level is because health plans are responsible for general medical and care for their populations and that's where they can work with all kinds of providers to do this. We don't intend this as a set of measures for psychiatrists alone or for family physicians alone. It's for the health plan to try to be looking at its population.

And then the data silos, they exist. We won't have any measures if we try to just focus on the data silos. And part of this is we're trying to get beyond the limits of the claims data where all we can do is look for tests. We can't look for a BMI, we can't look for an A1C result. We can't tell whether blood pressure is under control unless we try to confront those data silos?

CO-CHAIR BRISS: To kind of follow

specifically on the -- my question was going to be on the first part of that question. So it's the -- it strikes me that it's unarguable that these are the things that kill people with behavioral health conditions, right? And so it's unarguable that these are -- and they also kill a lot of other people, too, right? And that it's unarguable that there are treatment gaps, right, for people with behavioral health conditions.

But you could have elected to take kind of -- to kind of solve that problem in a variety of ways, right? And in some says this is a multiple kind of conditions problem and a measure parsimony problem, right? And so you've got perfectly good measures that address all these things in the general population, and it strikes me that you could have recommended that those measures be applied particularly to people with behavioral health conditions or recommended that the measures be stratified by people with

behavioral health conditions or a variety -- if you're HEDIS you could actually recommend that all their measures be applied in some year to -- well, to people with behavioral health conditions. And the advantage of that kind of thing could be that you don't wind up with -- for those of us who are generalists, I get a little bit -- there are probably dozens of kind of special populations for whom a hypertension measure could be applied. And they all sound -- taken one at a time they're all really reasonable until you wind up with 40 of these.

So can you talk a little about how you thought about that kind of stuff?

MS. HUDGSON SCHOLLE: Yes, and thank you.

You know, we talked with Helen about how we would implement these measures in a way that would make it not seem like we just created 11 new measures out of measures that already exist and trying to figure out how we

do it. But that's where we get back to the logic of can we demonstrate the high risk and the disparity? Because what we want is a standardized way to be able to manage and monitor the care of people with serious mental illness. If we don't have that standardized and measured in a way that people can report on it, then we're not going to get those data.

So that's the point of coming to NQF with these measures is to say, this is a high risk group of particular interest. It's particular interest in the duals work, it's of particular interest for Medicaid, it's of particular interest for Medicare. We're not going to get to that population unless we say, okay, we agree this is how we're going to define it and these are the measures we're going to use to monitor this high risk population.

So now whatever we can do to try to sort out the -- and make it sound like we're not -- that it has value for this population

and also to set up what would be the criteria for doing this again? So that's kind of a, how do you manage this in the measurement enterprise question? And you know, could everybody come up and say here's my favorite high risk population and I want to have a measure for them? Yes. Well then, let's set up some criteria for how you define what that is. And that's what we proposed here.

The other issue is, we started with 25 topic areas in behavioral health where we looked at importance and evidence and stakeholder support for areas, okay? This topic area was near the top because people were so concerned about mortality. And the evidence about the impact on outcomes is the strongest. And unfortunately some of the other things that we could do to try to improve the quality of care for people with alcohol and drug dependence or with serious mental illness, we're a long way from having the evidence about what to do and exactly how

to measure the quality. So this rose to the top both in terms of potential for increasing the life and quality of life for this population, and in the absence of other things that we really know how to do.

DR. BURSTIN: We actually spent a lot of time with Sarah and her team about this issue because it is a really important issue and it has implications not just for SMI but for many different populations. I think our thought was this is still the right approach for now. I think we would love to get this committee's insights to about how we maybe could make it more of a sort of matrix approach, these are the measures for all people and then there's a subset that apply to specific populations as needed.

I think that's work that PCPI and NQF really need to do together, but I don't think it should stop you from looking at these measures and moving forward. That's more of an issue of trying to sense make how these

come together, but I think the measures on their own still need to stand on their own and that's where they, you know, specifically target the highest evidence.

CO-CHAIR BRISS: And I'd like to see it, for that discussion, Helen, I'd love to see you explore this kind of thing as a stratification --

DR. BURSTIN: Absolutely.

CO-CHAIR BRISS: -- as opposed to creating lots and lots of new measures.

DR. BURSTIN: And some of that truly is our inability to create sub-measures under measures and truly just make that available and easy to see. We can probably link them and make it clear, you know, this measure connects to 1402 which is the general measure, things like that to just make it a little bit easier, maybe thinking about ways of pulling up the set and saying this is a population level set. But again, those -- that's work to be done.

CO-CHAIR BRISS: And I will now put my Chair's hat back on and go back to Jeff.

MEMBER SUSMAN: I think my comments were very much along this line of trying to develop a taxonomy that's robust. And right now it seems like we're just sitting all over the map in the taxonomy. One day we're talking about Peds, next day we're talking about SMI. Then we've got people with heart failure and, oh yeah, we'd better be measuring about depression occurrence in heart failure. Do you think it's time to take a step back and start saying, okay, whole population. Then we have different divisions some of which are by age, some of which are by disease, some of which may be other factors like racial or ethnic or socioeconomic breakdowns.

But rather than to go down this path which I think is a lot of work, I mean, I absolutely a hundred to a thousand percent condone the focus on improving, quote, medical health in patients with SMI. But I think this

approach is the wrong way to go. I think it's a lot of work for this group, a lot of measures we're going to consider here today, and I think there's a much more parsimonious way to do this work, efficient way, value driven way, and that would be to again create taxonomy in the ability to develop this interlinkage, the ability to cut populations in different ways which supposedly we're going to be doing already in looking at things like risk adjustment.

MEMBER TRANGLE: That means me?

CO-CHAIR BRISS: Yes.

MEMBER TRANGLE: I want to -- I agree with what you were saying about the taxonomy and the need to sort of somehow be able to look at a subset of the population as a disparity group. And then have resources really help you work on it.

We've been attempting to do this in Minnesota and have sort of a coalition or collaborative going where we're about five

years into it but where we took patients, SMI patients, and looked at our disparity on --

(Telephone ringing.)

DR. BURSTIN: Oh, this is probably just Obama.

(Laughter.)

MEMBER TRANGLE: Just Obama?

(Laughter.)

DR. BURSTIN: It is.

CO-CHAIR BRISS: Okay. I'll try again. Mike was talking to the mic. I know. I know.

DR. BURSTIN: It should be going by now.

MEMBER TRANGLE: God, horns, too, not just --

DR. BURSTIN: This happens pretty routinely when your people actually go to the White House. You get used to sort of --

MEMBER TRANGLE: Wow, that's a disparity group of one.

(Laughter.)

MEMBER TRANGLE: Anyway. So we've been trying to do this and we sort of replicated some of the national data using our public payers in the data, connected it with death and saw that our disparity was 24 years. Our main causes were cardiovascular which was 27 years of life lost, followed by accidents and injuries, pulmonary cancer.

We created sort of a bundle, we kind of likened it after the diabetes bundle and said, let's look at things separately but try and think of the idea like it was all or nothing like they do with diabetes, with the D-5, and came up with a bundle of an annual visit with a PCP measuring the BMI and -- and this is where I think I would like us to be thinking, forward thinking, because we can gnash our teeth and complain about what we can't get from claims until Obama gets replaced with four other presidents.

But what we did is said, BMI and what we wanted to be less than 30, you know?

Hypertension into the normal range. We had a lipid one but now that the standards have changed to where we're in total confusion about lipids, you know.

CO-CHAIR BRISS: As are the rest of us.

MEMBER TRANGLE: But we had hypertension and then we had either a hemoglobin A1C or a fasting blood sugar in the normal range. And then we did a high risk drinking or drug usage kind of screen to hearken to accidents and injuries which was up there, and we thought that was maybe the factor.

So we've been trying to work on this and we're about five years into it. And what we've found is systems with EMRs are able to sort of track this bundle and not get good data from people that aren't part of their systems. We're the land of mega-systems, you know, large integrated systems, so we can kind of do that. We've also found that we can

measure BMI, great, doesn't change, you know? But for things you can give a drug to, you know? So hypertension, lipids, you know, you'll measure it and actually see improvement. The drug screen doesn't get done as much.

We've been kind of working on this and it's all over the map. And to some extent, one of our biggest barriers to really doing a cohesive approach with any kind of real traction has to do with nobody -- the health plans and the other repositories of data do not subdivide their populations by SMI and have no way of doing that right now, you know? And aren't motivated to do that right now, you know?

To the extent that NQF and NCQA could think about at least piloting or testing how would it play out if somebody tried to look at this as a disparity group even though it isn't a hundred percent proven, you know, yet? It could be -- it could really start

getting us forward and I think this is one of those measures, like depression, and like we were talking offline about ADHD, that wouldn't it be nice someday as more and more of us are on EMRs that we actually do a symptom checklist -- whether it's a Connors' or Mackovac, whatever one you want for ADHD, and see whether people are getting better, not merely whether they're seen face-to-face within 30 days.

And I think this is the kind of -- this is the kind of thing that the only way we're really going to make progress besides creating a disparity group is to start doing it in places where we can also get data about what is going on with BMI and these other kind of submeasures, and not stick to thinking about where we're stuck with claims.

MEMBER ATKINS: I think about this in terms of if I were held accountable for this and at Medicaid ---- we have plans in 16 states. So the flexibility is ideal for us

because what I'm going to do in one market is going to be really different to achieve the same result than what I might do in another market. So I think that's actually required to have that degree of flexibility. Because a mature setting is so, so different ---- and what you're going to do about that, whether there's ACOs and so forth.

The issue -- and a few people have touched on this and this came up for every one of these measures so I'll put it out there now so we don't have to do it repeatedly. The business models that we have include fully integrated but we own all the data. In some settings that -- we're hoping will change over time. The plan has contracted with an MBHO where we have a delegation oversight and some degree of control over that, but it depends a lot on who you've contracted with. And then situations where the state carves it out, where we have no influence at all except through political informant stuff.

So ---- and I understand that in a theoretical world, the health plan really is accountable for everything that happens to its members. I would say to you that I don't live in that world and that the -- there's a very interesting sort of research question, not -- would each of these metrics actually, with some variation, be explainable based on those three different ways, care -- the plans are organized?

So I would say to you, it would be very helpful to be able to stratify the data. And I don't know that those are the three perfect models to stratify it against but to be able to look at -- you'll have better results when it's fully integrated and they're all my members versus the other two. And I think that's a research question that's sort of kind of a different question than the one you want to ask, but an important one for the industry.

MEMBER CHALK: I appreciate NCQA's

desire to split these up but I don't buy it.

(Laughter.)

MEMBER CHALK: I still -- as difficult as it is to make this -- to say this should be a stratified population health measure, I think it's important -- somehow or other, it's important either for this committee or some other committee or NQF to make that point. Because, once again, we're splitting out, yes, no, that ---- yes, patients with serious mental illnesses are at higher risk for diabetes. There's no question about that, there's plenty of research. But once again we're saying, oh, we're going to create now eight -- seven diabetes measures or eight diabetes measures for one population all split out and you wouldn't do that for -- you're going to do that for any other population or only SMI? And is -- are seriously mentally ill the only population that's -- subpopulation that's at higher risk? No.

And you know, I have my disagreements about -- there's no data about controlling high blood pressure related to people who have alcohol dependence, but that's a separate matter, I won't go into it. I really do have difficulty with saying to plans, even though I agree with your -- the problem that you're raising -- or states, you can cherry-pick these measures. You can pick which measure -- you know, pick one, implement that and so what? I don't understand what the so what is if you pick hemoglobin A1c and do nothing else. I mean, what -- so.

CHAIR BRISS: Caroline, you're up.

MEMBER DOEBBELING: Thank you.

I am concerned as well about this set of measures, not from -- in addition to all of the reasons that have been stated before, but one of the things that came up in the small group discussion about these measures were the performance of the measures over time. And I had raised the question

because these measures are and have been used in the general population for a period of time, yet none of that information was submitted. I came to learn during the course of that discussion that, because these are independent measures, they should be kept separate from that but nonetheless I still had a lot of curiosity about how each of these separate measures has performed over time since they have been used in the general population.

My concern about that is, if they haven't made much of an impact in the general population to get to a point where we think that they will start making a great impact in populations like the seriously mentally ill, for all of the reasons that you all have just been describing, I think is concerning and we need to think about that.

So I am not sure NCQA brought the data about past performance to the meeting today, but I do think we need to think about,

overall, how have these measures performed.

CHAIR PINCUS: I'm recused from talking about this measure specifically. But it just sounds to me like a lot of what we're talking about is -- almost relates to how one markets these measures or -- which I think goes to an issue more broadly -- which goes to an issue more broadly in terms of NQF's sort of way in which they catalog and -- you know, and utilize them. And how people who, you know, measure users utilize measures.

And is there -- I'm just wondering if there's a way to sort of separate that out? Because clearly there's work to be done by NQF to think about how measures get packaged and linked and you know, and it really -- and marketed, so to speak. Because in some ways, while you do an endorsement process, it also has elements of a marketing process.

DR. BURSTIN: Right, and it's marketing-based also.

CHAIR PINCUS: Right. And so that's

-- yeah, and so it might be useful -- I'm just suggesting that there might be a way to sort of separate that issue from the evaluation of each of these measures, per se.

DR. BURSTIN: Yeah. And in some ways I think it's -- I was just sidebar-ing with Peter, I think some of this is -- some of it's marketing but some of it's really about the implementation of how people actually use these. At least currently it's kind of beyond what we do, but we can certainly link measures on our database or something like that, just to make it clear that these measures hang together for the care of patients with SMI. It just hasn't been something -- and again I think it's a struggle for both NCQA and NQF, frankly.

You know, is it really a composite of the state, the individual? I mean, I think these will evolve over time and I guess we'd still want to make sure we're evaluating the measures on their merits that's before you

today. And a lot of these issues are really important policy issues we should continue to discuss but they shouldn't necessarily impact your evaluation of the individual measures.

CHAIR BRISS: And that's essentially where I was, too. So essentially what I was going to suggest is -- there seems to me to be a fair amount of feeling around the table that in addition to evaluating these measures individually, there's some additional work to be done sort of toward either composites or toward stratifying the parent measures or something. There seems to be a fair amount of sentiment around the table that we don't want to wind up having all the parent measures sort of spawn 40 sub-measures for every potentially important subpopulation. And so NQF and NCQA needs to think about that.

But I guess what I'd suggest now is that we go through the measures individually, unless people around the table feel so strongly about those issues that they want to

reject the set and have NCQA sort of go back and bring us back either a composite or a plan for stratification?

Michael?

MEMBER SUSMAN: I feel strongly but it's -- I think there is a missing player here. We're talking about NQF, NCQA, but there's a huge constituency of community mental health centers and public -- you know, where CMS and SAMHSA tend to drive what happens operationally. And so for example, one of our initiatives is every ACT team in our state has to report these measures, you know, and it's a different playground, and I would like them to be part of this discussion. I don't know how we get part of this, not now, but in the future.

MEMBER ROBINSON BEALE: You know, I think one of the things about these measures, and I think it was mentioned earlier but I want to re-emphasize it again, is that there's a lack of clarity across the country, across

health plans, across providers as to what HIPAA allows you to report and integrate and what you don't. And with that variability in that, I think it makes it really difficult to implement these measures.

So, for example, I sat with five different health plans who had different interpretations as to whether or not they could let their primary care physicians know that, A, their patient was an SMI patient, two, whether or not they were on a anti-psychotic, three, whether or not they needed diabetes tests. And so these are kind of fundamental, if you're going to hold a health plan accountable because they have all the data, but they can't do anything with the data because they're crossing over that magical medical/behavioral line without some clarity from HIPAA, that I think we've got a problem here.

And it's been an ongoing problem that I think -- and maybe those who are in the

room know better. I sense that they keep dancing around it but not necessarily making it exceptionally clear. I see a lot of confused faces. So if you've got a different understanding of this then I would greatly appreciate that.

CHAIR BRISS: Yes, sir, David.

MEMBER PATING: So I'm in favor of us moving through each indicator and then coming back and having a reconciliation discussion. I think what has to happen is, there's a whole -- so let's take the blood pressure measure. There's a whole general blood pressure measure, and how would this measure relate to that measure and can you -- you know, it's part of the stratification and subsetting that -- and it may be part of the marketing.

Once my system, for example, is on the hook for an elevated blood pressure, we have to track it by the primary care standards for blood pressure measure until it's done.

So there's things these things will plug into, but we should look at the individual measures now and then come back at the end and -- with a second set of stratification/reconciliation issues.

CHAIR BRISS: So I think that's right. So I'd like to -- unless anybody else has something that they urgently need to say, there are a couple of cards that are still up but I think they're left over, perhaps.

So what -- I'll give you the last word in a second, but what I'd like us to do is move to the individual measures next, unless somebody wants to move that we reject the set, which I didn't get much of a sense of the group of. And then just a reminder that -- just a reminder that we've got eight of these to get through in the next three hours, right? And so I'll need us to be pretty disciplined about not -- being crisp in your comments and not repeating things that have already been said, either in the general

session or in -- as we go through a lot of measures that relate to each other.

MEMBER ATKINS: So a couple observations. One with regard to why SMI? There are a couple populations of people in our country that have systematically been ignored, SMI being one of them. People with IDD being another. Kids in foster care being a third. We have to start somewhere. And so I think that the history of public sector behavioral health really drives this focus to try to reconnect with our public -- the folks who serve in the public sector.

With regard to the HIPAA issue, I think there's reality and delusion there. The reality is part two, which is not HIPAA, it's about substance use disorders specifically and that is a true and real problem -- I'm actually meeting with HHS next week to encourage them to do something about that.

HIPAA is internal counsel explaining to people why they don't have to do their

jobs, as far as I'm concerned. For behavior -- for mental there are no -- HIPAA does not preclude the coordination of information with PCPs and behavioral health treatment. It's an artificial issue, it's not real.

MEMBER SUSMAN: It does influence the marketplace actual practice.

MEMBER ROBINSON BEALE: There is confusion across very large national organizations. American Academy of Surgeons sat and talked to us about how they don't mix their behavioral health data with their medical data because of HIPAA. Well, I said, what does that mean? Well, we can't do it.

I'm just saying there's confusion. There needs to be clarity of the statement so that people can proceed.

CHAIR BRISS: So with that, let's try to move through the individual measures today. The first one is Body Mass Index Screening, so 2601. So Sarah, would you like to tee up the measure for us?

#2601: Body Mass Index Screening and Follow-Up For people with SMI

MS. HUDSON SCHOLLE: I thought I was free.

(Laughter.)

MS. HUDSON SCHOLLE: BMI, obesity, huge problem related to the medications that many people with serious mental illness are on, who is the most often topic that people told us we should develop a measure that looks at this as a sign, an early sign of metabolic problems.

This is based on the existing measure that is specified for the Physician Quality Reporting System, so it's a provider level measure that we used. But it is -- it looks at screening and follow-up. Now in this measure what we did is we looked at what was changing the denominator to address the serious mental illness population, and then changing the numerator because our panel said that a single event is not enough for people

with SMI where the PQRS measure is about whether there's a follow-up plan documented in the record at the time of the visit.

So physician reporting measures are based on the visit and we're looking at a health plan. And so we said health plans, you're responsible, not just for a plan being documented or something being done at the visit, you're responsible further, you know, for something happening. So a follow-up visit -- follow-up plan is not enough, there needs to be two events in the record that can be in the medical record or documented in other ways by the health plan in their care management systems. But those are the two main -- that's a change to this measure.

MALE PARTICIPANT: Within what period of time?

MS. HUDSON SCHOLLE: Within three months.

MALE PARTICIPANT: Two events in three months?

MS. HUDSON SCHOLLE: Right. The period of time for the follow-up, it's two events within three months of the BMI documentation, or the documentation of the higher score.

And actually what I did forget to tell you, which is sometimes confusing to people, the way these screening and follow-up measures, I think we're just doing -- the screening and follow-up measures, the logic of the measures are you meet the measure if you screen negative. So a BMI less than 30, which is not obese, right? Or if you are a BMI greater than 30 and you have the two follow-up events. So it's either -- meeting the measure is if it's not a problem based on the screen or the screen -- when you screen positive there are two events.

And the other thing I will note about the testing results is that we saw a huge variation in testing across the three plans that we looked at. The biggest reason

that people did not meet the measure criterion is because they did not have a visit at all with any kind of provider. And because we looked at -- all of the health plans we tested with had responsibility for both the medical and the behavioral health benefit. And I will agree that access to those behavioral health records are hard to get, regardless of whether you're responsible for it.

And so -- but in this case when we did have the records, it's not like we found a lot. We didn't find a lot of BMI testing happening in the behavioral health record but it was a problem to get those and that -- it becomes more of an issue for some of the other measures but not with this one.

CHAIR BRISS: So they're making me do double-duty as Chair so I also get to tee this one up. So I'm going to try to model brevity.

So this is clearly a high priority health condition, so we're -- so obesity is

common in the general population, probably commoner in this population. It obviously has significant health effects ---- so clearly a high priority health condition. Screening and follow-up is uncommon in the behavioral health population, not shockingly. And then the evidence on improvements in outcomes is there in a few good studies but the effects were frankly small. So it's there as -- sort of confirming the Minnesota experience that we just heard about, right?

So some very intensive interventions with up to 24 follow-ups resulted in about four percent declines in body weight. So if there were -- if I were to have a quibble with the intervention it's about the balance of the intensity of the interventions that have been studied and the health outcomes that have been shown.

No secondary reviewer on my list. Did somebody do a second review on this one?

MS. HUDSON SCHOLLE: No, for these

measures there was only one reviewer.

CHAIR BRISS: Okay.

MS. HUDSON SCHOLLE: Since they're all so similar.

CHAIR BRISS: So as the second reviewer, I agree with myself.

(Laughter.)

CHAIR BRISS: And would anyone like to -- with that, the table is open. Mike?

MEMBER LARDIERI: Thank you. I just have a question.

Why on the denominators is the -- for schizophrenia and bipolar, it's inpatient or two outpatient visits and for depression it's only an inpatient visit?

MS. HUDSON SCHOLLE: That denominator is consistent for all the measures.

The reason we did that is we were -- we looked at the research to see how the serious mental illness population has been defined, particularly in studies that have

looked at this high mortality risk. We queried our expert advisory panel about how we should do this.

What we were concerned about is that we didn't really have a definition of disability, right? It's hard to find that in the claims data, you can't find really disabled or chronic disease. And because depression is ---- can be an episodic, mild, recurrent -- I mean, mild condition, we felt that just looking through two visits with a depression diagnosis would kind of sweep in a lot of people that might not have serious mental illness that's disabling. And so that's why we followed the model that we found in the literature of using schizophrenia and bipolar wherever it exists as an inpatient diagnosis or two outpatient events, you know, just to reconfirm it wasn't an error.

And then for depression we said an inpatient event because getting hospitalized would indicate that it was a level of severity

---- we might not get at the chronic and disabling part, but at least a level of severity. So that's how we did it. And it did -- I mean, it narrowed down the number of people that got into the denominator when we applied that but we felt like that made more sense to us.

MEMBER ZIMA: I just want to add one other point, clinically, because that bias also kind of struck me -- I'm okay. And even though we might not be representing persons with major depression who are unable to access inpatient care, the other side of the coin is that, by excluding them we're less likely to have people on the atypical -- we're going to increase the risk that there's more people on atypical antipsychotics when we use this measure. And I think then that's where weight gain is more of an issue. And so in that case, I kind of felt like much more comfortable with that decision point.

MEMBER PATING: I just want a

clarification of the numerators. So you have EMI and then a follow-up care visit. I couldn't find the specs on the follow-up care visit. Does it have to be a follow-up care visit where BMI is coded or just any generic medication follow-up visit? Or maybe even not a medication visit, can it be a non-medical visit? So what constitutes follow-up specifically with regards to this measure?

MS. HUDSON SCHOLLE: There's a whole variety of activities that meet that follow-up criterion. And so we tried to model it on the existing measure which looked for counseling. And it could be counseling, nutrition visits. It could be pharmacotherapy. So I'm trying to find the page so I could tell you where it is, but it's a whole variety of services that are recommended by the U.S. Preventive Services Task Force and then also incorporated in the existing measure specification.

No, no, it has to be specific to follow-up.

MEMBER ATKINS: So with respect to that, this came up in our small group discussion. The first thing I noticed is the difference between this group and the general U.S., you know, task force is this SMI folks. I would think the first intervention is reevaluate the medication regime so that -- I mean, if you have an option to put them on a medicine that's less likely to cause weight gain, and ---- I mean, I've seen way too many of these come to me where they've been in treatment for years on meds and now they weight 350 pounds. So first thing I did was change them, you know? So I think that's a concern that I would have, that would be the first intervention I expect in behavioral health that's different than the general issue with obesity.

And the other one is around a list of meds and their -- and I saw something this morning, I heard it when I was like getting ready to come. Some medicine is on this

direct-to-consumer marketing about weight loss. There are several meds, other than the one that you list, that are weight-loss meds so I wouldn't want to limit it to that. I don't know how you'd do it but it's a moving target now, so I don't know how you would do it in respect. But you wouldn't want someone to lose credit if they didn't take their medicine.

MS. HUDSON SCHOLLE: So in terms of the medications, we did look carefully at the medications and we have a process for reviewing medications and adding medications. So there were two meds that were actually listed in the U.S. Preventative Services Task Force. One of those is not incorporated in the specs of the existing measure or our measure. And then we looked at the new measures that are coming out.

We actually have a process for updating measures when new medications come out, so we can respond to that. It's just we

can't do it like now because we have a process where we review it.

MEMBER ATKINS: I'm just saying --

MS. HUDSON SCHOLLE: Yeah. So we would do that, that would be part of our normal update, annual update of measures, to add that in.

MEMBER ATKINS: Okay.

MS. HUDSON SCHOLLE: In terms of the question about ---- we did wrestle with this question about could we look at change in medications and, you know, from the pharmacy claims, could you figure out that the meds were changed and then make a judgment that that was done to address the weight management? We felt like that was just kind of a little weird. I mean, it'd be hard to implement too. And so instead, that's why we thought, you know, the events of counseling -- and I do have to say the counseling on weight, if it's with the provider who did the screening or another provider, then what that

would mean is at that visit you said you changed it and then you saw them again about the BMI and that would count.

So we felt like actually the way this is set up that you get credit for that if you saw them and their BMI was high and you said, okay, I'm going to do it again, and we're going to change meds and do it. As long as you document that you addressed that weight management issue in the record then that would count.

MEMBER ATKINS: You might want to elaborate on that because at least I missed that and I was looking for it.

MS. HUDSON SCHOLLE: Okay.

CHAIR BRISS: I think -- Bonnie, is your card intended to be up? All right. So I don't think I see anybody else.

So the question is, are these -- do these specs mirror the ones from the general population measure?

MS. HUDSON SCHOLLE: And we did --

yes. The change was that we increased the number of events from a single event to a -- with two events within three months. And we did not allow ---- in the original measure, which is a provider level measure, you can say, referral to nutrition counseling. We don't -- our measure does not allow that, it's a health plan measure. If there's a referral, we want to see that nutrition counseling event for it to count.

CHAIR BRISS: So now I see no further cards up. And so we might be ready to vote on evidence.

MS. BAL: Okay. We're now ready to vote on evidence for 2601, and voting is now open.

We're missing one, if everybody could just try to vote again, please?

So the results for 2601 evidence is high, fourteen. Moderate, eight. Low, one. And we'll move forward to gap. And we're going to continue with the rest of the voting,

all of them together, correct? Okay.

Okay, voting is now open for performance gap. The options are one, high. Two, moderate. Three, low. Four, insufficient.

So we're only at 20 votes, if we could get everybody to vote again? Okay. Actually we've hit 23. We're good to go. Thank you.

Okay, so for performance gap for 2601 we have high, nineteen. Moderate, four, and we'll go forward with high priority. And the voting is now open.

Okay. If we could just have people vote one more time, I only have 21 and we need 23. Perfect. Thank you.

So the results for high priority for 2601 is high, seventeen. Moderate, four. Low, one. Insufficient, one, and we'll move forward.

CHAIR BRISS: So on reliability and validity, in general the workgroup thought that the measure was precisely specified and

clear in the fact that it's adapted from a currently implemented general population helps. The reliability was tested with having ---- with kappa scores from two graders and the kappa scores showed almost perfect in the rater reliability.

The clear specification and good agreement between raters also helps support a validity argument, and expert panel and public comments generally supported the face validity of the measure. There were -- the measures were tested in plans and showed low performance generally, and there were questions that we've already talked about, about ability to implement given data-sharing problems. But in general the workgroup seemed to feel reasonably good about reliability and validity of these measures.

And I'll concur with myself again. With that, the table is open. The floor is open for comments.

MEMBER PINDOLIA: I just had one

question on -- I can't remember the general population HEDIS score for this. What was that and how does this one compare to that?

I remember the diabetes and the -- I remember all of those, I just don't remember this one. I just want to know how big of a difference it was.

MS. HUDSON SCHOLLE: So the current HEDIS measure only looks at screening. And so we did compare it and these results -- there was a disparity, I mean these results are much lower than the current HEDIS measure when you just looked at the screening component. But this measure is looking for screening and follow-up.

MEMBER PINDOLIA: Right.

MS. HUDSON SCHOLLE: And a different NQF panel, I guess, recommended that we try to implement that measure in HEDIS and I don't know whether --

MEMBER PINDOLIA: No, I understand there's a difference but I was just trying to

figure out the part that is related. Was it 50 percent versus 30 or 80 versus 20 or --

MS. LIU: Ten percentage point difference between the screening rates.

MEMBER PINDOLIA: Thank you.

MEMBER MAZON JEFFERS: I have a question. So the difference between this measure and the BMI measure for the general population for HEDIS is, first, that it includes a follow-up component and, second, that the denominator includes a definition of SMI that's based on the definition that you -- that is here, right? So there are really two differences to the measure?

MS. LIU: That difference is only comparing the screening rate. So you know, the material that didn't include the screening rate in the final measure rate.

MEMBER MAZON JEFFERS: So I guess I would just say that those seem like pretty big differences.

MEMBER CHALK: Am I right that this

was tested only in Medicaid plans and this is only to be -- no?

MS. HUDSON SCHOLLE: It was tested in a Medicaid plan that was for disabled adults.

MEMBER CHALK: Right.

MS. HUDSON SCHOLLE: A Medicaid plan for low-income adults, no disabled, and a dual SNP.

MEMBER CHALK: Yeah, that's what I thought.

MS. HUDSON SCHOLLE: So it's a special needs plan but that's a Medicare/Medicaid.

MEMBER CHALK: Yeah, right. So it's three public sector plans. So it's not -- NCQA is not going to put this out to be used by commercial health plans, right? It's only a public sector measure?

MS. HUDSON SCHOLLE: Well, that's up to NCQA to determine with its --

(Inaudible comments.)

MS. HUDSON SCHOLLE: This gets to the implementation of the measure. How it's going to be used is really different.

CHAIR BRISS: Yeah, as a general rule, Mady, the answer is once you approve a measure it can be used by anyone for anything. That's right.

So any other comments before we move to voting? All right, let's vote.

MS. BAL: Okay. Voting is now open for reliability.

Oh, the options are one, high. Two, moderate. Three, low. Four, insufficient.

Okay. The results are high, ten. Moderate, nine. Low, four. Insufficient, zero, for reliability of 2601. And we'll move forward to validity now.

And voting is now open. The options again are one, high. Two, moderate. Three, low. Four, insufficient, and we're voting on validity of this measure.

Okay. The results for validity for

2601 is high, ten. Moderate, eight. Low, three. Insufficient, two. And we'll move forward to discussion on feasibility.

CHAIR BRISS: So the subgroup had, given that these sorts of measures are currently implemented in the general population, that these measures were likely feasible. There was some discussion about chart abstraction being a burden -- some burden, but that doesn't make it infeasible. And so the committee had few feasibility concerns. And with that I will open the floor to general discussion.

It appears to me that we can move straight to voting.

MS. BAL: So voting for feasibility for 2601 is now open. And the options are one, high. Two, moderate. Three, low. Four, insufficient.

MS. DORIAN: Caroline, are you on the phone? Were you planning to vote on this?

MEMBER DOEBBELING: I am. I forgot

to hit the send button.

MS. DORIAN: Okay. Just checking.

(Laughter.)

MS. BAL: So the results for feasibility --

MEMBER DOEBBELING: My typing didn't just automatically transmit. Sorry about that.

(Laughter.)

MS. DORIAN: No problem.

MS. BAL: The results for 2601 feasibility is high, nine. Moderate, seven. Low six. Insufficient, one. And we will move forward with this measure and start discussion on usability and use.

CHAIR BRISS: So in general, the committee discussion about -- again this is based on a currently in-play population measure which generally supports usability. The stakeholder review was generally positive. There were issues raised that have come up around this table about -- and it seems to me

to have been at least partially dealt with about the kinds of interventions that might be considered.

But in general, I think the committee discussion had raised relatively few issues. And with that, the floor is open for any additional comment.

We vote.

MS. BAL: Okay. Voting for usability and use for 2601 is now open. Options are one, high. Two, moderate. Three, low. Four, insufficient information.

So the results for 2601 usability and use is high, five. Moderate, thirteen. Low, four. Insufficient information, four -- I'm sorry, one. Four percent, but one. And so we will now vote on overall suitability unless there's further discussion.

CHAIR BRISS: Anybody want to make any closing arguments to the jury before we give the overall vote?

Oh, a hanging judge, who knew? All

right. Let's vote.

MS. BAL: Okay. Voting is now open. The options are one, yes; two, no.

We do need one more vote in the room, if we could get everybody to just hit one more time?

Perfect. Thank you. So final result is yes, twenty, no, three, for 2601. So this measure is being recommended.

CHAIR BRISS: Excellent. So we've picked up a little bit of time, we now have seven measures to do in two and a half hours.

(Laughter.)

CHAIR BRISS: And even though that was efficient, I don't recommend to the staff that they make anybody else do a one-person show of Chair primary and secondary reviewer again. So thank you.

If NCQA would tee up the hypertension measure for us? Oh, and just for discussion, I'm anticipating that there are going to be all kinds of issues that come up

with these measures that are sort of repeats of things that we've already talked about. So it's perfectly fine, I think, to say we've already talked about this issue in previous measures so that we don't have to spend a lot of time on it again, if it's just a repeat.

#2602: Controlling High Blood Pressure for People with SMI

MS. HUDSON SCHOLLE: Okay. So this measure is applying the controlling high blood pressure measure for people with serious mental illness. It uses the same denominator. We focused on blood pressure because of the high risk of cardiovascular disease in people with serious mental illness due to lifestyle factors, side effects of treatment and disparities in care.

And so the only thing I might point out -- it's tested in the same group and we saw disparities. Our stakeholders -- we conducted focus groups with stakeholders that did include folks from community mental health

centers and other -- and consumers and states and a whole variety of potential stakeholders. This one and the diabetes results, the level of disparities that we showed, I think people said -- we asked were these results surprising, did they make sense? They said, yes, not surprising but heartbreaking to see the poor level of care.

And the measure specification reflects the new specifications that NCQA has put out for the 2015 measure specifications for this measure, which will have an age -- different blood pressure expectations depending on age. We tested it using the measure specifications that were consistent with the 2012 reporting so we could make comparisons. But the specs are aligned with that existing measure.

CHAIR PINCUS: So actually just to clarify, Sarah, because I didn't get to do it on each one of these, to say if they're exact -- if the specifications for the numerator are

exactly the same as the one for the, you know, referent measure?

MS. HUDSON SCHOLLE: Right. So for the controlling high blood pressure, the only change to the measure is to the denominator, of narrowing the denominator ---- or specifying it for serious mental illness. The numerator is exactly the same.

CHAIR BRISS: So I have Caroline as kicking off our discussion for the committee, please.

MEMBER DOEBBELING: Thanks, Peter.

This measure is a measure that, during our small group conversation about it, we found to be -- echoing Sarah's comments, an important measure because of the discrepancies between the SMI population and the general population with regard to measuring and controlling blood pressure. And also the significant morbidity and mortality as related to hypertension.

I don't really have any more

comments to make other than we did find it to be an important measure.

CHAIR BRISS: So ---- Raquel?

MEMBER MAZON JEFFERS: I just had a question. Why are pregnant women excluded from the denominator?

(Inaudible comments.)

MEMBER MAZON JEFFERS: But it's a blood pressure. It was also excluded from the body mass, so I thought maybe BMI might be related to pregnancy. But blood pressure --

(Laughter.)

CHAIR BRISS: Probably has something to do with pregnancy, right?

MS. LIU: I mean, it's consistent with the HEDIS general population measures exclusion because pregnancy would affect their blood pressure. So that's the reason.

MS. HUDSON SCHOLLE: So it's excluded from the BMI because we're looking at weight and so that's why it's excluded from that one.

For the hypertension measure, it has to do with, would you expect doctors work to reduce the blood pressure in pregnant women within the same timeframe that you would for the general population of people with hypertension, right? Because we're looking for a diagnosis of depression. It probably also -- you can't figure it out during that year, right, because you're -- in this measure you're identifying people who have hypertension from the claims. You're confirming the diagnosis in the medical record in the first six months of the year and you're looking to make sure that the last blood pressure of the year is under -- is meeting your threshold. And so that's why pregnancy would make it complex to implement.

CHAIR BRISS: Other questions or comments before we move to voting?

MEMBER ZUN: I noted in the measure there is a comment taken during an acute inpatient stay or ED visit. So I'm not sure

I understand this. So who's obligated then to ensure that the patient gets connected with their primary care doctor?

MS. HUDSON SCHOLLE: So what we're looking for are people that have a diagnosis of hypertension that's confirmed in an outpatient setting. Because of concerns about white coat hypertension or hypertension that might be picked up in an ED visit only, then those visits are excluded because that might not be a real diagnosis of hypertension. That's what the measure is getting at. So we are looking for people that have hypertension. And remember, with every measure what we want to do is we want to make sure that we're finding the right people. And so sometimes that means you exclude people that ought to be in the denominator but you're trying to go for specificity rather than sensitivity. So that's an explanation for that.

CHAIR BRISS: And isn't it also generally true in all of these measures that,

because you're specifying at the plan level or higher that you're asking that people get the right care, you're not micromanaging how the hand-offs get done. Isn't that right?

So with that, any other questions or comments before we move on to voting?

MS. BAL: Okay. Voting for evidence for 2602 is now opened. Options are one, high. Two, moderate. Three, low. Four, insufficient evidence. Five, insufficient evidence with exception.

MS. BAL: Okay. The results for evidence for 2602 is high fifteen, moderate seven, low one, insufficient zero, insufficient evidence with exception zero. And we'll move on to gap. And the voting is now open. The options are one high, two moderate, three low, four insufficient.

Okay. The results for performance gap for 2602 is high sixteen, moderate six, low one, insufficient zero. And we'll move to high priority. And the voting is now open.

Same options, one high, two moderate, three low, four insufficient.

Okay. The results for high priority for 2602 is high eighteen, moderate five, low zero, insufficient zero. And we can discuss reliability and validity now.

CO-CHAIR BRISS: So Caroline, can you tee up reliability and validity for us, please?

MEMBER DOEBBELING: I sure will. The comments from the workgroup on reliability and validity were very similar to those brought up in the earlier discussions. The measure is described well and the group felt that it was precisely specified and clear. And given that it had already been implemented in the general population, we understand that the measure works for the population. There were no significant concerns about the specs themselves.

The concerns in this area were brought up largely about whether or not the

health plans reliably could have some data that were going to be measured for fragmentation and care resulting in a mixed picture of what we're really seeing with the data. And all those things that were mentioned earlier about behavioral health carve-outs, questions regarding HIPAA, and that type of concern. The stakeholders generally supported the face validity of the measure.

There was concern that the small sample size did not provide sufficient data to conduct statistical tests, and there was still a comfort level with the fact that the data did suggest meaningful differences across plans with the general population. And I had my question about how well these measures have performed in the general population over time, and that was not addressed.

CO-CHAIR BRISS: Sarah, would you like to comment on the general population performance?

MS. HUDSON SCHOLLE: Right. So the measure has undergone some changes in specification that make it a little bit hard to look at it. But I would say that over the past five years, we've actually not seen much improvement in the measure at the health plan level for Medicaid plans. I think we see more improvement in some of the other plans, where this measure is actually being used in other kinds of pay for performance arrangements.

CO-CHAIR BRISS: Any other comments or questions around the reliability and validity? Hearing none, let's vote.

MS. BAL: Okay. Voting is now open for reliability. The options are one high, two moderate, three low, four insufficient. And this is for 2602. We only have 20 votes so if everybody could just try to do it again? Thank you.

And the results are, for reliability for 2602, is high nine, moderate seven, low six, insufficient four. And we'll move

forward to validity. And the voting is now open. Again, the options are one high, two moderate, three low, four insufficient. We need one more vote from the room. Please make sure to vote once the timer is up. If the timer's not up the vote won't be registered. Thank you.

So, for validity we have high nine, moderate eight, low four, insufficient two. And that will be enough to move us forward to discussion of feasibility.

CO-CHAIR BRISS: Caroline, you're on again.

MEMBER DOEBBELING: Thank you. The workgroup's comments on feasibility were around two issues. One, the requirement for medical record abstraction, which creates a burden on plans, especially for plans in which the SMI are enrolled in the health plan, but receiving their care in a very fragmented system about where and how to find that chart information. It didn't mean that these

measures are not feasible, but only difficult to get to.

And then the same concern about feasibility, with regard to the overall fragmentation of care and behavioral health carve-outs was brought up again during discussion. Some aspects of the measure can be captured from electronic sources, but not all are well maintained in an electronic sources. Overall, the feasibility discussion was much like the other measures, and no significant concerns were noted.

CO-CHAIR BRISS: Any comments from the floor?

MEMBER PINDOLIA: So on this one, I guess I do share that concern too, because in the state of Michigan there just is not any coming to the same EMR interface for behavioral health and the physical health for our dual eligibles. And I think other states may be facing that. Is there something -- exceptions, is there something available for

health plans that cannot get information fed back of what the behavioral health component has found?

MS. HUDSON SCHOLLE: You're really getting into an implementation issue about how these measures would be used. So the -- these are feasible, these are imminently feasible, if people allow you access to the data. And while I don't believe that there are any laws that preclude that, it's really a matter of will and a force of will strong enough to overcome people's concerns about it and the challenges of doing it.

And remember, for a health plan that's -- if the health plan were responsible for this, they could make it happen. A health plan that's responsible for the medical care of this population of people with SMI, they can make this measure happen by making sure that the doctors that they're paying do this, even if -- in the general medical care, even if it's not happening in the behavioral health

setting. That might be a duplication of services. But they've got the ability to do that, and that's part of what we're trying to force -- we're trying to encourage them, is to pay enough attention to this.

But it's an implementation. So if the state of Michigan said these measures don't work, we're not going to -- we don't think it's feasible, then that would be how they might deal with it.

CO-CHAIR BRISS: Any other comments or questions before we move to voting? Hearing none, let's vote.

MS. BAL: Okay. Voting for feasibility is now open. The options are one high, two moderate, three low, four insufficient. I'm assuming Larry didn't vote, so we're just going to go ahead.

Okay. So for feasibility for 2602 we have high seven, moderate nine, low five, insufficient one. And we'll move forward to usability and use.

CO-CHAIR BRISS: Caroline, one more time.

MEMBER DOEBBELING: Sorry. The workgroup had no significant comments about usability other than those that have already been mentioned. Nothing new to add to that discussion.

CO-CHAIR BRISS: I love that summary. Thank you. Any questions for -- any questions or comments from the room? Let's vote.

MS. BAL: Okay. Voting is now open for usability and use. We have quite a few votes still out there, if everybody could please vote. We should be looking for 22.

Okay. The results are -- for usability and use for 2602 are high six, moderate eleven, low six, insufficient zero. And we can go forward to the overall decision unless we have discussion.

CO-CHAIR BRISS: Any closing remarks? Hearing none, let's vote.

MS. BAL: Okay. Voting is now open, overall suitability. The options are one, yes, two, no.

Okay. The final result is yes, eighteen, no, five, for 2602. So this measure will be recommended.

CO-CHAIR BRISS: Terrific. Thank you. So, I have good news and bad news. The good news is that we're moving very fast, the bad news is that we just finished our morning, right? So with that, 2603, Sarah, you want to tee that one up for us, please?

#2603: Diabetes Care for People with SMI: Hemoglobin A1c (HbA1c) Testing (NCQA)

MS. HUDSON SCHOLLE: Okay. So NCQA calls the diabetes set of measures a measure, and they call each of the items within the set indicators. So this is one of, I think, six indicators that we're -- or measures that we're bringing to you. And all of these measures are similar to the blood pressure in that what we have done in our specifications

is we have defined people with SMI as the denominator. All of the numerator statements for this entire -- for the remaining, all the diabetes measures are the same. So it's the same set of numerators. So we haven't made any changes in those at all.

And I think the testing results were pretty consistent with what we found in the controlling high blood pressure. Again, lower performance rates. Interestingly, I would point out that the performance rates were very different for the different plans, and so the -- the plan that served low-income adults had the poorest performance rate, and that's because even among people with diabetes, very few of them had visits. And so not having a visit will contribute -- if you don't have something, then you'll automatically fail it.

For the -- the plan that did the best was the dual SNP. And remember that many of these measures are included in the Medicare Stars Program. And so there are special

incentives for Medicare plans.

But in addition, the dual SNP actually has -- is set up as a system to try to find people, and they have other responsibilities for managing care for this population.

CO-CHAIR BRISS: So with that I have Lisa Shea as the --

MEMBER SHEA: Yes, thank you. So in brief, our group thought that this was a very important measure, that there was demonstrated gaps. And that's all I'll say.

CO-CHAIR BRISS: I may have created a monster.

(Laughter.)

CO-CHAIR BRISS: Okay. Comments from the room? Let's move straight to voting.

MS. BAL: Okay. Voting for evidence is now open for 2603. Just confirming, everybody in the room has voted? Did someone walk out I'm not aware of?

If everybody could just vote one

more -- oh, we got it. Thank you. Okay. So for evidence for 2603 we have high nineteen, moderate four, low zero, insufficient zero, insufficient with exception zero. So we'll move forward to the gap vote. And it is now open. Still short one, so everybody please make sure to vote.

Okay. Perfect, thank you. So for gap we have high twenty-one, moderate two, for 2603. And we can move forward to high priority. And the voting is now open.

Okay. The results for high priority for 2603 is high nineteen, moderate four, low zero, insufficient zero. And we can move forward to scientific acceptability.

MEMBER SHEA: So, again, in terms of the reliability, the workgroup generally felt that the measure was precisely specified and clearly bolstered by that it's already used in the general population. And that -- I'm looking at the reliability. And they had really high inter-rater reliability results,

kappa was very high.

Then regarding validity, they -- the measures were tested in the three plans and there was a lot of variability in the performance, as we've heard. And while there wasn't sufficient data to conduct a proper test, the group felt that there were meaningful differences that were likely to exist, and felt overall comfortable that there was validity, in terms of this measure.

CO-CHAIR BRISS: Comments from the room? So I think we can move to vote. Oh, I'm sorry.

MEMBER PATING: I'm just wondering, and I'm not an expert on diabetes, just the age or going down to 18 with HbA1c, was there other ways that that could have been framed? You know, your risk for diabetes goes up perhaps with age, with BMI, and I just don't routinely test my 18-year-olds for A1c. So I was just wondering, does --

MEMBER SUSMAN: This is SMI -- this

is SMI population.

MEMBER PATING: Oh, they have -- oh, I apologize.

CO-CHAIR BRISS: I think we can move to vote.

MS. BAL: Okay. Voting for reliability is now opened for 2603.

CO-CHAIR BRISS: We still seem to be missing --

MS. BAL: We're missing -- yeah. We're missing one in the room. Okay. The result for reliability for 2603 is high sixteen, moderate five, low two, insufficient zero. And we'll move forward to validity. And the voting is now open. We actually need one more from the room so everybody could please make sure that they voted.

Okay. Perfect. Thank you. And so for validity of 2603 we have high fourteen, moderate five, low three, insufficient one. And we can move forward to discussion of feasibility.

MEMBER SHEA: So like the other measures, it was deemed to be feasible. And one point I guess I would make is that the more the plans adopt these measure, then the less burden will be on them because they'll be in the same chart looking at the different measures.

CO-CHAIR BRISS: Excellent. So any comments from the room? Hearing none, let's vote, please.

MS. BAL: Okay. Voting for feasibility is now open. And just to remind you, please make sure to point at me.

CO-CHAIR BRISS: Not at me, at her.

(Laughter.)

MS. BAL: Okay. The result for feasibility for 2603 -- right? Yes -- is high ten, moderate nine, low four, insufficient zero. And we can move forward to usability and use.

MEMBER SHEA: So usability, basically the same concerns or issues that

came up before. But our workgroup, in general, felt that this was a usable measure and generally favored its use.

CO-CHAIR BRISS: Any further comments or questions? Hearing none, let's vote.

MS. BAL: Okay. Voting is open for usability and use. So actually we -- I thought we were waiting on the phone. But if everybody could retry, we're missing one in the room. Just in time.

(Laughter.)

All right. So for the usability and use for 2603 we have high thirteen, moderate six, low four, insufficient zero. And we can move to overall suitability unless there's discussion.

CO-CHAIR BRISS: Any final discussion before we vote? Hearing none, let's vote.

MS. BAL: Okay. It is now open to vote. The options are one, yes, two, no.

Okay. The result for overall suitability for 2603 is yes, twenty-one, no, two. So this will be recommended. And we can move on to the next measure.

CO-CHAIR BRISS: So let's go ahead and, we're about at our proposed break time, so let's do take our break, 15 minutes, and restart at 25 after, please. And we'll still have five measures to do in about 80 minutes then, so let's do -- be reseated and ready to go at 25 after, please.

(Whereupon, the above-entitled matter went off the record at 3:04 p.m. and resumed at 3:25 p.m.)

CO-CHAIR BRISS: Can we get restarted, please?

(Inaudible comments.)

CO-CHAIR BRISS: So the next one is 2604, the diabetes care.

#2604: Diabetes Care for People with SMI: Medical Attention for Nephropathy (NCQA)

CO-CHAIR BRISS: I'm sorry, it's

nephropathy.

MS. HUDSON SCHOLLE: Nephropathy, same story, same population, nephropathy --

CO-CHAIR BRISS: Same song, different verse?

(Laughter.)

MS. HUDSON SCHOLLE: Right. Right. Just this is about identifying the screening for one of the major complications of diabetes. And where there's concern, and other evidence that people with serious mental illness don't get this screen.

CO-CHAIR BRISS: So Bob, will you tee this up?

MEMBER ATKINS: So I won't just say ditto, but I will say that we agree that both diabetes and nephropathy are bad and we should do what we can to make them less bad, yes.

(Laughter.)

MEMBER ATKINS: There's no -- and so the group entirely agreed that this was -- that there was adequate evidence, more than

adequate evidence to support this as a focus, that -- sticking with the plan -- that there is clearly a performance gap, and this is very high priority, that it's high risk, high cost, problem prone and it's a terrible thing for it to happen to people. And we need to do something about it. So that's the summary.

CO-CHAIR BRISS: So I can hardly wait to go back and explain to my family that what I did with all these experts today is determine that bad is bad and better is better, right?

(Laughter.)

And does anybody else have comments about the evidence for this one? Let's vote, please.

MS. BAL: Okay. Voting for evidence is now open. Just making sure that we have enough people in the room. So yes, for evidence -- for a quorum, we have enough people to vote. And we do, no worries at all.

For evidence we have high fifteen,

moderate five, low zero, insufficient zero, insufficient with exception zero. And so we can move forward to gap. And voting for gap is now open.

Okay. So for gap we have high nineteen, moderate two, low zero, insufficient zero, and we can start voting on priority in one second. We can start voting now.

Okay. The results for priority is high sixteen, moderate six, low zero, insufficient zero. And we can discuss scientific acceptability now.

MEMBER ATKINS: Okay. With regard to reliability and validity, there are really minimal concerns and they replicate those that have already been spoken to. With regard to both reliability and validity -- I just want to make sure there's nothing from the group as a whole. No, we've already talked about all the issues.

CO-CHAIR BRISS: Anybody have comments before we vote? Hearing none.

MS. BAL: Voting for reliability is now open.

Okay. The results for reliability for 2604 is high fourteen, moderate five, low three, insufficient zero. And voting for validity is now open.

Okay. The results for validity for 2604 is high eleven, moderate seven, low four, insufficient zero. And we'll move forward to feasibility discussion.

CO-CHAIR BRISS: Back to you, Bob.

MEMBER ATKINS: Feasibility. Again, there was really no difference from the comments made on the prior measure. We're already doing this, so there's no real increase in burden. And same issues concerning getting the data.

CO-CHAIR BRISS: Anybody want to comment before we vote? Hearing none.

MS. BAL: Voting is now open for feasibility. We're just waiting on one more vote, if people could just revote, please?

Thank you.

Okay. So the final vote for feasibility is high twelve, moderate eight, low two, insufficient zero. And we can move forward to discussion of usability and use.

CO-CHAIR BRISS: Bob, anything new to add on --

MEMBER ATKINS: There's nothing to add about usability and use. It really is identical.

CO-CHAIR BRISS: Would anybody else like to find something new to add? Hearing none, let's vote.

MS. BAL: Okay. Voting is now open for usability and use. We're missing two votes -- okay, one vote. Please just make sure to point at me.

Okay. The result for usability and use is high ten, moderate nine, low three, and this is for 10 -- I'm sorry, 2604. And now we can move to forward overall suitability unless there's further discussion.

CO-CHAIR BRISS: As always, I'll give you a chance to make a closing argument if you'd like. Hearing none.

MS. BAL: Okay. Voting is now open for 2604, overall suitability. Options are one, yes; two, no.

Okay. The final result is yes, twenty-one, no, one. So this measure will be moved forward for recommendation.

CO-CHAIR BRISS: So with that done, the blood pressure control and diabetes measure?

#2606: Diabetes Care for People with SMI: Blood Pressure Control (<140/90 Malmstrom Hg) (NCQA)

MS. HUDSON SCHOLLE: Okay. So it may be a little bit confusing that there is a blood pressure control measure that's part of the diabetes set, and then there's the blood pressure control measure that we already talked about. But let me try to explain the Venn diagram here, because there could be some

overlap, but not necessarily.

So the controlling high blood pressure measure focuses on people with hypertension, okay? And it looks to confirm the hypertension diagnosis and then -- within the year, and to see that blood pressures so controlled. The blood pressure measure that's for diabetics is for everybody who has diabetes, regardless of whether they carry the hypertension diagnosis. So some people are going to be in that -- they're going to have both hypertension and diabetes as their diagnoses and they would show up in both samples. But some people would not, and so that's why there are two separate measures.

MEMBER TRANGLE: That's still within SMI, right?

MS. HUDSON SCHOLLE: Right. This is -- right. I believe that there's some evidence that blood pressure is the thing you want to control for people with diabetes, like it's a very important indicator for people

with diabetes.

CO-CHAIR BRISS: Yes. So before we get into the details of this measure, I'd actually like to -- let's have a little bit of a harmonization discussion. So it feels to me like these are mostly overlapping on one -- I actually have some questions about whether you really need the second diabetes measure, the diabetes and hypertension measure.

If you've already got the hypertension measure, you could just apply it to the population with diabetes. It feels like this is mostly a historical artifact of a time when we used to treat to different targets in diabetes and hypertension. Bob, do you want to -- you can just pull up a chair.

MR. REHM: I do this at the office all the time. I kneel before our NCQA Gods. So I'll genuflect later.

(Laughter.)

MR. REHM: My hands aren't clasped. So the -- just for historical context, the --

and I'll just use the NQF as a frame. Currently there's an NQF endorsed measure for blood pressure control for people with hypertension, that's essentially the corollary of what we just talked about. We also have an NQF endorsed measure about blood pressure control for people with diabetes.

So in terms of lineage, these are both NQF endorsed and, as I recall, you know, strongly endorsed. There was not a lot of disharmony about that. They were aware that the blood pressure measures existed in each other's space and I think they, at least those panels, respectively, felt that these were appropriate delineations. And then Sarah's point about the Venn diagram is accurate.

The hypertension measure that you've just reviewed is somewhat unique because it is very, very focused on a confirmed, this confirmation of a hypertension diagnosis for all the reasons that Sarah alluded to. The diabetes measure is really, you know, do you

have SMI with a comorbidity of diabetes, and then do you have your blood pressure control? In some ways it's a simpler measure, it's just looking for that one value, the one value being the latest value and it's a little bit more straightforward.

I know it may seem like parsing but, Peter, I think that the market appreciates it. From the health plan perspective they manage patients with diabetes, and hopefully someday soon, patients with SMI as holistically, and think about them, and think about the things that they can intervene with -- distinct maybe from another population. And that's helpful to them. Sometimes they do it, sometimes they do it different ways, but at least it's a tool in their toolbox.

MEMBER SUSMAN: I guess I'm still a little bit unclear since we have an SMI measure with hypertension and diabetes and now the hypertension guideline is basically set at the same specification what the added value is

here. I still haven't heard that. Maybe I'm missing it, I'm sorry if I'm delaying our progress.

MR. REHM: So I'll try to be helpful. The -- what's unique about is the time element. The blood pressure measure you just, you know, recommended for endorsement, is looking at a -- trying to capture essentially a -- I wouldn't call it a new hypertension diagnosis, but a confirmed within a confined period of time. And then giving the health plan through its provides, but giving the health plan, you know, essentially -- call it six months. It can be broader than that. But enough time to engage that patient and bring their hypertension under control.

The diabetes measure is essentially, once I'm on that diabetes denominator, you know, and assuming the diabetes isn't resolved because they've lowered their BMI, and good things have happened in their life. But that population is just going to continue to

persist and show up every year, year in, year out, with a little bit of fallout. And then you're just seeing, is their blood pressure controlled along with several other indicators, like nephrology consult or any of the other indicators that are in A1c, testing and control. So it's just one of many. But it's a slightly different frame.

MS. HUDSON SCHOLLE: So the real unique part that the diabetes measure will get, that's people with diabetes, regardless of whether they have a diagnosis. So it will be people whose diagnosis isn't confirmed in the record. It's every person with diabetes, regardless of whether they've got a defined -- whether they've ever carried that hypertension diagnosis.

CO-CHAIR BRISS: It's true, there are lots of people with hypertension that are, my boss would say hiding in plain sight, right? So there are lots of people who don't care, who probably have hypertension who don't

carry a diagnosis. The flip side of that is that there are a lot of people who, at a point in time, can be over 140, or over 90 that don't actually have hypertension.

MR. REHM: And the way we structured the blood pressure measure, controlling blood pressure measure is such that the clinician or a health plan who observed that high rate has time to go back and go back. It's not just one reading, it's the opportunity to, if you will, escalate. And it's essentially the last reported blood pressure of that period, wherever the measurement period lands, it's the one that's used to see whether he met the threshold. So it's a -- it's giving people both time and encouragement, and incentive to do something as opposed to not doing anything at all.

MEMBER TRANGLE: This is interesting. And one of the distinctions I seem to be hearing is that if one has diabetes the thought is you have a chronic disease and

they're going to measure this indefinitely. And if you just have hypertension it may go away. If you think about the pool that we're talking about here, which is SMI patients, I think the approach towards these patients in general is that they have a chronic disease, and we need to kind of continue to monitor them over their lifetimes to see how they deal with their BMIs and other kinds of risk factors.

So I get the distinctions of the timing, I'm just sort of, as I think it through, I think what we might be evolving to is, if we do start thinking about SMI as a sort of disparity group with their own things that we want to be monitoring for their lifetimes, it would be more like -- it would be diabetics, and within our cluster we just look at this. We're not there yet but I think that would be the evolution.

CO-CHAIR BRISS: I'm still fuzzy now about what constitutes control in the two

groups. I get that the diabetes measure might be more inclusive of people with elevated blood pressures, because it doesn't require people to also carry a hypertension diagnosis. But I don't get what -- I don't get what the difference is in terms of the measure of success on the measure, right?

MR. REHM: So one frame to think about it up here is that the diabetes measure with the threshold is -- it's saying I come in on one -- I'll just use a number, 130 over 39 or -- I mean, 138 over 79, pardon me. And then I'm encouraged to say, boy, that's cutting it close, you know, I could monitor, I could do a variety of things even though I meet the standard of the measure.

There's a quality improvement component to this, it's not just -- I know we talked about accountability, but there is something about this, its function is much early warning and helping people stay below the line, if you will as much as it is finding

those that are above the line who you want to bring down below. So it's operating at two -- I guess two different levels.

CO-CHAIR BRISS: But success on the measure, so separate --

MS. HUDSON SCHOLLE: So among people with SMI and diabetes, we're looking for their last blood pressure of the year, regardless of whether they had a hypertension diagnosis or whether that's the first one in the year, we want it to be below the threshold.

For people with hypertension, we say first we have to identify you as being hypertensive and then we're going to give you six months to get to the lower rate. So the idea there is that the -- they're just different populations. And they're going to be overlapping, but there will be some people that are different in each group.

MEMBER PATING: So could I ask, regardless of SMI status, so you have those that are SMI and we're checking for diabetes

and SMI we're checking for blood pressure. What about those with diabetes, are you required to check for blood pressure? Because I really want to make sure that we're staying consistent with the general population, the measures.

MS. HUDSON SCHOLLE: Yes. This is exactly the way it's done in the general population.

CO-CHAIR BRISS: So, in some ways you don't have to think of anything new, because you're a primary care clinician. Because you have the diabetes algorithm, your SMI algorithm and your blood pressure algorithm. All right, I'm sorry I derailed this a little bit but this was a little too easy for NCQA. So we needed to have a little cross-examination, right? And so --

MEMBER PATING: Why can't this just be a sub-measure or is it being considered a sub-measure? To me it's like it doesn't quite rise to its own status. It should be linked

to one of the other measures, part B, or something like that. Those are my thoughts.

MS. HUDSON SCHOLLE: They're all kind of sub-measures in what we're trying to do is just take the logic of saying this is what you do for diabetes and apply it to people with SMI. Same thing people with hypertension, you apply the same logic. And the original measures have that overlap, and that has been acceptable to the field. And so we're not questioning it.

CO-CHAIR BRISS: Acceptable to parts of the field. So Rhonda, I think you were going to be -- yeah, you were going to be the lead discussant on this before I short-circuited the discussion. Are there other things you'd like to add as we move through?

MEMBER ROBINSON BEALE: Essentially, I don't have anything more to add to this one, unless there's others who would like to add comments to this.

CO-CHAIR BRISS: So let's try

voting.

MS. BAL: Okay. So the vote for evidence is now open.

So the vote for evidence for 2606 is high fifteen, moderate five, low three, insufficient evidence zero, insufficient evidence with exception zero. And we'll vote on gap now. Gap is open.

Okay. The results for 2606 gap is high sixteen, moderate six, low one, insufficient zero. And we'll move forward to vote on high priority. And the vote is open now.

The vote for 2606 high priority is high thirteen, moderate five, low five, insufficient zero. And we'll move forward to discuss scientific acceptability.

CO-CHAIR BRISS: Rhonda, do you want to add anything on this one?

MEMBER ROBINSON BEALE: I'm a group of one. There were no -- I don't have any other comments to make on this.

CO-CHAIR BRISS: Anybody else, questions, comments or concerns?

MEMBER ROBINSON BEALE: I think it's all been said.

CO-CHAIR BRISS: All right. Hearing none, let's open the vote.

MS. BAL: Okay. Voting is now open for reliability for 2606. Could everyone just vote one more time? We're one person short. Thank you.

Okay. So for reliability we have -- wait, validity. The score is high thirteen, moderate eight, low two. And that was for reliability for 2606. And now the voting for validity is open. So just need one more vote. I don't know if maybe someone stepped away, so we'll just go forward with it.

So for validity of 2606 we have high eight, moderate twelve, low three, insufficient zero. And we can move forward to perhaps discuss feasibility.

CO-CHAIR BRISS: Anybody have

anything new? Rhonda, have anything?

MEMBER ROBINSON BEALE: Nothing new on that one, unless someone else has anything.

CO-CHAIR BRISS: Anybody want to add anything? CO-CHAIR BRISS: Let's vote.

MEMBER ROBINSON BEALE: Anything new in the discussion? No? Great.

MS. BAL: Feasibility is now open for voting.

Okay. The result for feasibility for 2606 is high seven, moderate thirteen, low three, insufficient zero. And we can move forward to usability and use.

CO-CHAIR BRISS: Anybody have anything new? Hearing none, let's vote.

MS. BAL: Okay. Voting is now open for usability and use. It did take me a second, so make sure that you pushed it after the timer came on.

Okay. The final result for usability and use for 2606 is high seven, moderate eleven, low five, insufficient zero.

And we can move forward to the overall vote, unless there's some discussion.

CO-CHAIR BRISS: Seeing no moves toward discussion, let's vote.

MS. BAL: Okay. Voting is now open. One is yes, two is no for overall suitability for endorsement.

Okay. So for overall suitability for 2606 we have seventeen yes and six no. So this measure will move forward for endorsement. And we can move on to 2607.

CO-CHAIR BRISS: NCQA will kick this off. Sarah?

#2607: Diabetes Care for People with SMI
Hemoglobin A1c (HbA1c)

MS. HUDSON SCHOLLE: I just want to point out that there are two measures that look at A1c control. One looks at poor control, so that's an A1c that's greater than nine. And the other looks at good control, A1c less than eight. And I think -- so basically the rationale for these measures is

the same. And the reason for doing this is that there's good agreement that greater than nine -- that nobody should be above nine. And but there is considerations about how far you should go in getting to good control. And so that's why we've had the greater than nine measure for a long time, but less than eight hasn't been in as long. But it's helpful to understand where your population fits. That's why there are two.

MEMBER SIDDIQI: So okay, did you want me to go ahead? I was going to say I will echo what Lisa had said. Evidence shows that it's -- that it's important, it's a major risk for morbidity and mortality and essentially measures the quality of care that we provide to diabetics with SMI. And there's, you know, evidence that there's disparity as to how those people are managed.

So do you want me to say more?

CO-CHAIR BRISS: Not unless you feel like there's something else new to say.

Would anybody else like to comment?

(No response)

CO-CHAIR BRISS: Let's vote.

MS. BAL: Okay. Voting for evidence for 2607 is now open.

(Pause)

MS. BAL: So we're still missing two in the room, if everybody could just vote. Let's make sure no one stepped out.

(Pause)

MS. BAL: Okay. The result for 2607 evidence has a high nineteen, moderate four, low zero, insufficient zero, insufficient with exception zero. And gap is now open for voting.

(Pause)

MS. BAL: We just need two more votes, if everybody could please vote. Thank you.

(Pause)

MS. BAL: Okay. So 2607 gap is high eighteen, moderate five, low zero,

insufficient zero. And now we can vote for high priority.

(Pause)

MS. BAL: So we only have 17 -- or 19 now. Please make sure that everybody's voting. We should have twenty-two in the room and then one on the phone.

(Pause)

MS. BAL: Okay. So we have -- for high priority we have high sixteen, moderate six for 2607, and we can start discussing scientific acceptability.

MEMBER SIDDIQI: Reliability, again it seems to be specified and clear. I don't know if anybody has any concerns that they're not reliable measures.

(No response)

CO-CHAIR BRISS: So let's move to voting.

MS. BAL: Okay. Voting is now open.

(Pause)

MS. BAL: Okay. So for reliability

for 2607 we have high thirteen, moderate eight, low two, insufficient zero. And now we can start voting for validity.

(Pause)

CO-CHAIR BRISS: We may have a comment on this one. Pay no attention to the scores behind the curtain.

Yes?

MEMBER MARK: I had a comment, question about the validity. So what do the guidelines say that the appropriate population based A1c levels should be? And do we have any concern about, you know, getting people too low and, you know, causing iatrogenic hypotension?

CO-CHAIR BRISS: Or hypoglycemia?

MEMBER MARK: Yeah, hypoglycemia, I guess. Yeah. So not being a clinician, but would --

MEMBER SIDDIQI: Long-term effects --

CO-CHAIR BRISS: We're still in the

greater than nine, but the poor control.

MEMBER SIDDIQI: Yeah. The poor control.

MEMBER MARK: So it's not an issue at all?

MEMBER SIDDIQI: It's not an issue at all.

MS. HUDSON SCHOLLE: I don't think there's any concern that people with diabetes should have A1c's less than nine. The concern -- oh wait, I'm sorry. Or less than eight.

The concern has been in the lower range, and there's -- where NCQA actually does report a measure less than seven which is really marked for quality improvement. But that measure, we did not present that measure for this group. That's where the concern has been, going less than that. So I think there's very good agreement on these two thresholds.

CO-CHAIR BRISS: And some of this is -- some of that concern is why there are two

thresholds represented in these measures, right? Most everybody agrees that nobody essentially should be greater than nine and probably the less than eight one might not be sort of a hundred percent.

(Pause)

CO-CHAIR BRISS: Are there comments on the reliability or validity?

(No response)

MS. BAL: Okay. Voting is now open.

(Pause)

MS. BAL: Okay. For validity of 2607 we have high ten, moderate ten, low three. And we can move forward to discussion of feasibility.

MEMBER SIDDIQI: Any concerns or any questions regarding feasibility?

(No response)

CO-CHAIR BRISS: Anybody else?

(No response)

CO-CHAIR BRISS: Let's vote, please.

MS. BAL: Okay. Voting is now open.

(Pause)

MS. BAL: Okay. The result for feasibility for 2607 is high ten, moderate ten, low three, insufficient zero. And we can discuss usability and use now.

MEMBER SIDDIQI: All right. Usability, I think we all agree based on discussion that it's very useful.

CO-CHAIR BRISS: So if we have no further ado, let's vote, please.

MS. BAL: Okay. Voting is now open.

(Pause)

MS. BAL: Okay. So for usability and use the -- for 2607, the results are high eleven, moderate seven, low four, insufficient zero. And we can move to overall suitability, unless there's further discussion.

(No response)

MS. BAL: I'm going to take that as a "no" and voting is now open.

(Pause)

MS. BAL: Okay. The final result is

yes, twenty-one, no, one. And this measure is being moved forward for recommendation. And we can move forward to 2608.

#2608: Diabetes Care for People with SMI: Hemoglobin A1c (HbA1c)

CO-CHAIR BRISS: And we have probably surfaced anything in the discussion of 2607 that needs to be said on 2608. Does anybody from NCQA or, you know, the -- or anybody else like to talk more about this measure before we just vote it through?

(No response)

CO-CHAIR BRISS: Why don't we try to vote it through, please.

MS. BAL: Okay, perfect. So voting for evidence for 2608 is now open.

(Pause)

MS. BAL: Okay. The result for evidence for 2608 is high nineteen, moderate three, low zero, insufficient zero, insufficient with exception zero. Voting for gap is now open.

(Pause)

MS. BAL: Okay. So for gap for 2608 we have high eighteen, moderate five. Priority's now open.

(Pause)

MS. BAL: Okay. So we have high seventeen, moderate five, low zero, insufficient zero for gap of 2608 -- I'm sorry, that was high priority for 2608.

And now reliability for 2608 is now open.

(Pause)

MS. BAL: Okay. So for reliability for 2608 we have high fifteen, moderate six, low two, insufficient zero. And now voting for validity is open.

(Pause)

MS. BAL: So we have for 2608 reliability, we have high ten, moderate eight, low four, insufficient zero. And that was for validity, my mistake.

Feasibility, the voting is open now.

(Pause)

MS. BAL: Okay. So we have feasibility for 2608 high eleven, moderate eight, low four, and we can move forward to use and usability. Voting now open.

(Pause)

MS. BAL: Okay. So we have for usability and use, we have high eleven, moderate six, low five, insufficient zero for 2608. And just a reminder for overall suitability, that one is yes, two is no. And voting is now open.

(Pause)

MS. BAL: We're at 21 so if everybody could just please vote for this one?

(Pause)

MS. BAL: Okay. So for the overall suitability we have twenty yes, two no and 2608 will be moved forward for recommendation. And we can move on to 2609.

CO-CHAIR BRISS: Sarah, would you like to tee this one up? Anything special

about this one?

#2609: Diabetes Care for People with SMI: Eye Exam (NCQA)

MS. HUDSON SCHOLLE: Other than this is a really big disparity. I think the -- it's an even bigger disparity than the others between the test plans and for SMI population compared to the others. So even poor access to this specialty.

CO-CHAIR BRISS: So Caroline.

MS. DORIAN: Is Caroline on the phone?

CO-CHAIR BRISS: Caroline, are you still there?

MEMBER DOEBBELING: Yeah, I'm still there. I'm in the car.

I concur with the (telephonic interference) the gap here is tremendous. I think it's driven in large part by today's(telephonic interference) primary care referral to specialty care for this, and it's a barrier to get the SMI into our specialty

care exams. So the gap is (telephonic interference) and the group was in agreement about all of that. For all of the other reasons, that we will discuss: validity, reliability, usability, feasibility, everything (telephonic interference).

CO-CHAIR BRISS: Caroline, that was an amazingly cogent discussion while driving and picking up your kids. That's very good.

MEMBER DOEBBELING: I am in the parking lot behind about 50 cars right now, so it --

(Laughter)

CO-CHAIR BRISS: Okay. So with that, I think we can move to the voting, please.

MS. BAL: Okay. So evidence for 2609 is now open.

(Pause)

MS. BAL: We are at 21. If you could just all vote again, please? I'm just trying to get as many of you here.

(Pause)

MS. BAL: Okay. For evidence of 2609 we have high nineteen, moderate three, low zero, insufficient zero, insufficient with exception zero. And performance gap is now open for voting.

(Pause)

MS. BAL: So performance gap for 2608 -- I'm sorry, 09, is high eighteen, moderate four, low zero, insufficient zero. And we can move forward to high priority.

(Pause)

MS. BAL: So for high priority for 2609 we have high fifteen, moderate seven, low zero, insufficient zero. And now we're going to move on to reliability and voting is now open.

MEMBER PATING: Open? I was wondering if I could just ask a question first. Oh sorry --

(Laughter)

CO-CHAIR BRISS: Somebody has to

occasionally do that just to make sure that we're all awake.

MEMBER PATING: If I could just ask the developer the specifications for the eye exam, you know, I guess how you chart it. If your eye exam is done in the inpatient setting, as part of, you know, the routine admission or if the psychiatrist does the eye exam in the office? I mean, you've just to find ways, I think, to have the eye exam.

MS. HUDSON SCHOLLE: An eye exam is -- it's a specialty eye exam.

DR. BURSTIN: It's got to be a dilated eye exam so it can't be done just in a regular office of a non-eyecare professional.

(Pause)

MS. BAL: So are we ready to vote?

Okay. Voting for reliability is now open.

(Pause)

MS. BAL: We are only at 18 so

please be sure to vote.

(Pause)

MS. BAL: All right. So for reliability for 2609 we have high fourteen, moderate seven, low one, insufficient zero. Voting validity is now open.

(Pause)

MS. BAL: We are at 19 so just one more time, please?

(Pause)

MS. BAL: Okay. So for validity of 2609 we have high twelve, moderate seven, low four, insufficient zero. And we'll move forward to feasibility unless there's discussion?

(No response)

MS. BAL: Seeing none, voting is now open.

(Pause)

MS. BAL: Okay. So feasibility for 2609 is high eight, moderate eleven, low three, insufficient zero. And now we can move

forward to usability and use. Voting is now open.

(Pause)

MS. BAL: Okay. So for usability and use for 2609, high nine, moderate ten, low three, insufficient zero. And we're ready for overall suitability. The options are one, yes; two, no. And we are now open for voting.

(Pause)

MS. BAL: So the final result for 2609 is yes, twenty, no, three, and this measure will be moved forward for endorsement. And I can no longer say this today. We're done.

(Laughter)

MS. BAL: Until tomorrow.

NQF Member and Public Comment

CO-CHAIR BRISS: So our last task for the day is to listen to public comments. So operator, if you could open the phone lines for public comment for us, please?

OPERATOR: At this time, if you would

like to make a comment, please press * and the number one on your keypad.

And there are no public comments at this time.

CO-CHAIR BRISS: There do not appear to be public comments in the room. So I'd like to thank everybody for a hard day's work, and we had a very efficient afternoon and actually finished a little early.

Yeah, we wanted to loop back, and so we've now worked through that new set of measures one at a time and we were going to loop back and see if anybody had any further thoughts about how they fit together. I mean, we talked at the beginning about -- there was at least some support in the committee for moving toward either composites or for moving toward a stratification discussion that might allow us to also capture other high need or high risk populations without creating hundreds of measures. Does anybody else have wisdom to impart after the specific

discussions today?

Yes.

MEMBER ROBINSON BEALE: This is actually more of a question. I think we all are kind of feeling that there are a lot of measures and there's a sense of overload. One of the questions I have is whether or not, for NQF, is part of your work to -- how do you say -- prioritize measures as their feasibility or usability, as it relates to things that are being done now with some of the measures? Some of them are just measuring an outcome or a baseline or population measures. And then some of them are starting to quickly get tied to performance, payment and other kinds of things. Does NCQA take -- does National Quality Forum take on the role of prioritizing or assessing the readiness of measures for those different venues?

DR. BURSTIN: Funny you should ask. Yes, that's become a very hot topic these days within NQF and other circles. In fact we have

a meeting next week about Consensus Task Force Group I mentioned to you, and we are actually going to ask them to help us make a decision of whether we should, in fact, move forward of endorsement that either is related to intended use of a measure or potentially even a rating system for measures where you can rate it based on the quality of how well they do against those criteria, as well as whether they've already been in use. So those are the two options.

We'll likely then do an extra panel and try to figure out the how to make it happen. But those are really important questions. And I think this is also, you know, part of what Sarah said, some of this also gets into the implementation space of how -- you know, how NCQA might prioritize or group these in terms of HEDIS. But I think it's also clearly, you know, in their bailiwick. But I think there is an important discussion.

And if any of these are part of any of the federal programs, they would also likely come up as part of the Measures Applications Partnership for that.

MEMBER TRANGLE: You know, my thoughts are exactly along those lines. It's like the nature of these things and the nature of just how progress happens in medicine is they're all going to metastasize and spawn, it will be like rabbits, you know?

And we have to, I think, think about not just prioritizing but how can you, instead of like subdividing it so that they're pure and simple and discrete, how do you integrate them? And how do you actually think about simplifying and looking at the overall burden? And when is the bang worth the buck in terms of actually doing it at the clinic, you know, or hospital, whatever it is level?

And if you don't have a process for that, my recommendation is maybe you're starting with this little thing about how do

we harmonize them so they don't conflict? But I think it's got to integrate and simplify no more than it's -- no slower than it grows.

MEMBER SUSMAN: I think that we really should be moving to the concept of perfect care within disease entities like diabetes. Take five measures, put them together, is it really acceptable to hit sixty percent or better, or seventy percent on each one of those individually but only deliver perfect care two percent of the time? I think obviously we're further away from that with individuals with SMI but we really should be driving to consistent, reliable care for a disease entity. And that's where I think we can start developing these composites that make sense, they have a lot of shared attributes in interacting ways where controlled blood pressure, lipids, your diabetes, A1c, all that stuff, weight, whatever, fits together.

MEMBER CHALK: Going back to the

issue of the MAP, I had a question about whether at some point the measure developers or the measure implementers should be required to talk about, given what we've heard today especially, action plans. And that that should not be -- whether that should be part of their submission of a measure. We heard a lot today and we'll get more tomorrow about measures that don't move. And there are reasons measures don't move, Sarah mentioned a few, I could mention some more.

To just have NQF be engaged in endorsing measures that go nowhere, even if they're implemented, because there's no action plan -- and I don't -- you know, health plans have action plans. They can put performance incentives in place. But that's not the only possibility, there are all kinds of possibilities. To help plans and states and other purchasers focus on improving the quality of care.

And it's very discouraging to me

that we're going through the exercise of approving measures and saying nothing about the action that needs to follow. We're not requiring developers who submit to the NQF, especially, which is the gold standard, to have an action plan as part of their submission.

CO-CHAIR BRISS: So I endorse the idea of more all-or-nothing composites. As I've said already, I would much rather have recommended stratifications and -- for high priority populations like the SMI population, but not limited to the SMI population, as opposed to having different measures for every population of interest which I think reduces the signal to noise and reduces progress in the main. And I think that there's still more work to do on harmonizing measures.

I know I didn't get anywhere with the hypertension measures today but, you know, at HHS we -- you heard me tell the story, in 2010 we worked on harmonizing measures and we

were working on hypertension, among other things. And there were some number like 35 different hypertension related measures. And all of them taken individually had something that somebody thought was a good rationale that sort of, in a vacuum, Bob could undoubtedly explain to me. But taken together, you know, it was just an awful mess and so much more can and should be done.

And those are my thoughts.

CO-CHAIR PINCUS: I do think that -- and it's just a thought in terms of some of the -- again, some of the feedback at NQF. One is, I'm in developing some more specific templates. So we talked earlier today, for example, on the patient-reported measures that clearly distinguish how to handle that as a performance measure rather than as a clinical instrument. And so that there's a different way in which one can sort of lay that out. And you know, perhaps it's not by breaking it necessarily into four different measures, but

to think of it as a domain of outcomes that's sort of utilized in different ways. So to think of ways of simplifying that.

Secondly, I think another kind of template that could be developed is just what you were talking about before is a template for thinking about high priority populations segmentation that could be a kind of complementary-type review that wouldn't require a full review like we've gone through every single measure this afternoon. But that could be a way to do that, you know, for existing measures in a much more simplified kind of process.

And then I think -- you know, I mean, I think what you're getting at, maybe it might fit with, in a way with what Helen had alluded to in terms of the fit for purpose efforts. But this whole issue of marketing or, you know, implementation or whatever you want to call it, about getting measures, meaningful measures out there to be used, and

that the strategy for which -- what kind of use this measure will have so it's not just flinging out there where it's on a list.

The more that that's thought through in terms of, well, who -- you know, who is this being addressed and for what type of program should be part of the assessment that's done for endorsements, because I think that that's -- so that -- and also listening to the context of, you know, how many other measures that are -- not just in -- you know, it's sort of a matrix. Measures aren't particularly a domain, but also measures for a particular use and sort of -- you know, just look at the cells in that matrix.

MEMBER JENSEN: I would just like to make a comment as the individual here from the Veteran's Health Administration. Certainly our recent issues could be attributed to a number of different causes. But I think it's not incorrect to say that some of the issues in Veteran's Health Administration could be

related to the vast number of performance measures that we were being held accountable to. And some pressure on senior leadership at facilities to demonstrate needing those performance measures and perhaps putting pressure inappropriately on providers or perhaps even not being quite honest with reporting measures.

So I just think it's very important to think about what measures do we really need to look at that are indicating that we're giving good care to our population, for me, the veterans, but our citizens? And not push people into being so concerned about their performance rating and their salary based on that that they're going to be dishonest and take measures that are inappropriate.

MEMBER PINDOLIA: So in regard to the meeting and, Helen, that you discussed that NQF is having next week, and maybe this is being done through public address and I'm missing it. But it seems like the measures

that are out there for three or six years and coming back, there is feedback, I know, provided directly. It's solicited by CMS for Five Star measures annually about what are some concerns that you've having about it. And then I don't think I ever get questions from NCQA, if you're having any reasons or rationale of why -- you know, if there's concerns on those measures.

But I think it would be very important for NQF to hear directly if a measure is staying static for three or six years, what is the feedback from the individuals that are actually doing the measuring, the health plans or other providers, of what's holding that back? And I'll give you osteo as a classic example. So we were told to do chart review. We did the chart review for 300 people and it's just 85 percent false positive in how it's being coded. So do we really have a quality problem? We don't know. We're using it off

of a miscoding, but it continues to be a measure. That's just one example.

And so when we talk about tomorrow, where I was part of that group, that was why I kept bringing up every time when the measure shows no improvement for three years, it just -- I don't know if that -- if it's really improving quality by continuing it in the same way.

MS. HUDSON SCHOLLE: We do -- NCQA does have a policy clarification system where we get comments all the time. They're reviewed every -- when we reevaluate a measure, we do a public comment where we ask people what's going on, if we're having problems. And so we do see that. And then -- and we try to address any of those concerns. And that's why you see some of the really -- what look like really detailed things and you wonder, why is that in the measures? Because somebody's made a complaint or asked about it and we've gone back and said, okay, we have

addressed that particular consideration, yes.

And there are measures, some measures can do some things but they can't do everything. And so I do want to recognize that, that it really depends on how they're used as well. And --

MEMBER PINDOLIA: And I agree with that. So I -- if you don't mind, I just want to just comment back.

I agree that if it's staying stagnant and it's just because that metric can't do everything, but then maybe through NQF, there should be some surrogate measures that can help increase it. I know the whole onus is it's supposed to be outcome driven. We say you're supposed to achieve this, now provider, health plan, whoever, you go figure it out. But if after three to six years we can see they're not figuring it out, maybe there should be like, okay, let's take one step back. Just because our whole goal is

really to get the quality to improve. And I just -- and that's what I struggle with, when they're asked for 100 or 52 measures and then they're like going berserk.

MS. LIU: So I think, you know, when we take the national average of several hundreds of plans, it's hard to see a huge movement from year to year. But if you drill down to original levels you'll see larger improvement. And also observe large gap between -- among plans, potential low performing, high performing plans, there are, you know, 20 percent to 40 percent difference between those plans. So we try to, you know, have that as a tool for plans to monitor and how they can be better.

DR. BURSTIN: Just to respond to Vanita, thanks for adding that in. I think that's such an important issue broadly across the measurement enterprise. We just -- we don't have very robust feedback loops about what's happening on the front line, what's

working, what's not. There's lots of pockets of it that, you know, very well develop developers like NCQA and Joint Commission do. But overall, we just don't know very much about which measures move the needle and which measures don't. And I think that's -- and what are the factors that go with it? You know, people often point to the, you know, the significant reduction in the early elective deliveries, for example -- apologies to the psychiatrists -- you don't have much time to spend thinking about obstetrics.

But you know, there's reasons why that measure went down. You know, it's a standardized measure, there was complete agreement on the part of all the stakeholders that this was bad for moms, bad for babies. I mean, so there's a whole series of things, and I just think we need to increasingly think about -- and some of this gets back to Mady's point about how this all comes together to create sort of the plan in the grouping, or

back to your point about how to take the outcome and sort of nest it perhaps with the process measures. It's just about sense making, I think, and there's been very little sense making in our measurement system. So we'd love to take on a bigger role there.

MEMBER PATING: In that respect, I would be interested, I mean, if it's possible on some measures, coming back at like five year reviews, particularly as you go through new coding systems and just looking at -- you know, given where we are now, would we have launched some of the measures were did, you know, in cycle one or zero, and these things evolve. And so I know that it's done at the NCQA level and the shop level. But in terms of the National Quality Form, giving it sort of a cushion. We launch ships but we don't see them coming back in some way.

DR. BURSTIN: I mean, it's actually important to remember. I mean, at least probably about a year ago we did the analysis

and a 100 new measures came in and 100 measures went out. So there is an effort to make sure that when there are measures that have just outlived their time, they need to go, but that's hard to do because there's a lot of people who are pretty tied to those measures for lots of programmatic reasons and other reasons.

So you know, I think we just want to make sure that whatever people are collecting and analyzing is actually helping to drive quality. But I think that's going to be a dream.

Well I mean, you do have, as part of all the maintenance measures, you know, what's in there for use and usability should improve the trends, it should improve where the measure is currently. And actually Peter and Harold are part of a standing committee focus group we did with the chairs of a whole group of our standing committees who made some pretty strong recommendations about how to

change our three-year maintenance process, being much more heavily oriented towards these issues of use and usability.

CO-CHAIR BRISS: Now we've -- the good news is that we've kind of reached a steady state, and the bad news is that the total of the claim measures is still something like 700.

DR. BURSTIN: No, it's 600.

(Laughter)

CO-CHAIR BRISS: I consider that to be a rounding error, Helen.

DR. BURSTIN: It's across many different settings and compilations.

CO-CHAIR BRISS: All right. So does Caroline want to get back in?

MS. DOEBBELING: I would if I could. Thanks, Peter.

In listening to the conversation, I think that there are a couple of things that we haven't brought up. I believe in the measures and what we are trying to do, but I

often am voting no because practically, I have seen the National Quality Forum endorsement being used against providers or health plans to say this National Quality Forum approved, therefore it's fantastic and therefore you have to use it. But the practicality of doing that is something that I think we don't talk about and we don't -- from the platform that NQF has, to help prioritize the measures or to say that for all primary care providers, these are the top ten things that will drive health in our population. It's weight, it's tobacco, it's blood pressure, it's X, Y or Z. And that's what we need to focus on.

From a very practical point of view, have a health plan working with providers, we have significant, significant access issues. The marketplace has only made those access issues harder because there are now more insured patients vying for smaller numbers of slots to meet many of the requirements. So the members fall into some plan's denominator,

but practically there is no place to put that person.

I don't think that the measures are going to help us change the access issues. I think that's something entirely different and a different type of conversation that we need to bring. But what happens then is providers and health plans become penalized by that because they have patients in their denominators that they practically can't be in their clinics because of those issues.

I also think the other practical issue at the provider level is that any provider office who has a mix of patients is going to have a mix of insurers. Whether that's Medicare, Medicaid or any of the private insurers. And so what I have seen is, to the extent that health plan with the richest incentive program or the richest pay for performance program, that's the plan that gets the most attention from the clinic. And so in Medicaid, I don't have the funds to

compete against some of the lucrative pay for performance programs or commercial plans, especially when the commercial plans may make up that provider's vast majority of their population. If it's a huge practice, you know, 80 percent versus 20 percent or something.

And so in populations where we don't see the needle being moved, I think that we have to think about what those reasons are for that. And I think the NQF would be in an ideal position to start taking a look at those broader issues.

CO-CHAIR BRISS: So this has continued to be a great conversation. I do just want to remind us that we're nearing the end of our time and it might be great to continue this conversation over dinner and perhaps a glass of wine or so. And within the bounds of healthy alcoholic use, right?

So Michael, do you want to have the last word?

MEMBER LARDIERI: I don't know if it's good to have the last word here.

But I'm listening to the conversation and I agree with what Caroline was saying. And I think some of the proliferation of the measures, like what we saw today, the measures for behavioral health are pretty much the same as they are for any other patient medically. And we might be able to lower the number of measures but require that they get reported on across race, ethnicity and then different populations. It's the same measure, and I just want to see -- I have some discomfort with behavioral health being pulled out as something separate.

We're trying to integrate, as opposed to keep us separate, so it should be the same measure. But then we should measure across race, ethnicity. I don't think we can not do that. And then across, okay, SMI and then maybe other populations. Instead of coming with a new measure for a specific

population, use the same measure and just say, okay, now we have this special population that we want you to measure on and it's all the same.

Because at the provider level, I agree with Caroline that, you know, if you have ten different plans you have ten different measures, it's really difficult to do that at the provider level. But if you have one measure that you just stratify and report on across race, ethnicity and sub-populations, I think that might reduce measures, get everybody on the same page, be easier to report on in the future.

CO-CHAIR BRISS: That seems like a great last word. Anybody else want to add?

Any more comments from the staff about -- anybody need reminders about dinner or where that happens or those kind of issues?

MEMBER DOEBBELING: And I'll look forward to my glass of wine.

(Laughter)

CO-CHAIR BRISS: All right, Caroline.

MEMBER SAMPSEL: So dinner reservations are at 6:00 p.m. at Mio. Mio is just north of L on Vermont, so out of the hotel, go up to L, take a right, take -- go two blocks, take a left and it's right there. It's on your left. It's eleven-something-or-other. You really can't miss it.

CO-CHAIR BRISS: So if anybody would like, I'll figure out an address by then and we can meet in the hotel lobby like a quarter of 6:00 and walk over. Is that okay?

MS. BAL: Kathy, could you just let everyone on the line know that we're adjourning for the day and we'll see them in the morning?

(The meeting was adjourned at 4:47 p.m.)

L

C E R T I F I C A T E

This is to certify that the foregoing transcript

In the matter of: Behavioral Health Phase 3
Standing Committee Meeting

Before: NQF

Date: 10-01-14

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