NATIONAL QUALITY FORUM

+ + + + +

COST AND RESOURCE USE STANDING COMMITTEE

+ + + + +

WEDNESDAY
MARCH 15, 2017

+ + + + +

The Standing Committee met at the National Quality Forum, 9th Floor Conference Room, 1030 15th Street, N.W., Washington, D.C., at 9:00 a.m., Brent Asplin and Cheryl Damberg, Co-Chairs, presiding.

PRESENT:

BRENT ASPLIN, MD, MPH, Co-Chair
CHERYL DAMBERG, PhD, Co-Chair; Principal Senior Researcher, RAND Distinguished Chair in Healthcare Payment Policy
LARRY BECKER, Retired*
MARY ANN CLARK, MHA, Vice President, Avalere*
JENNIFER EAMES HUFF, MPH, CPEH, Principal, JEH Health Consulting; Senior Advisor, Pacific Business Group on Health
TROY FIESINGER, MD, FAAFP, Physician, Chairman Quality Committee, Village Family Practice of Fort Bend*
NANCY GARRETT, PhD, Chief Analytics Officer, Hennepin County Medical Center
ANDREA GELZER, MD, MS, FACP, Senior Vice President and Corporate Chief Medical Officer, AmeriHealth Caritas*
MARTIN MARCINIAK, MPP, PhD, Vice President, US Medical Affairs, Customer Engagement, Value, Evidence and Outcomes,

GlaxoSmithKline

KRISTINE MARTIN ANDERSON, MBA, Executive Vice President, Booz Allen Hamilton
JAMES NAESSENS, ScD, MPH, Professor, Mayo Clinic
JACK NEEDLEMAN, PhD, Professor of Health Policy and Management, UCLA Fielding School of Public Health
JANIS ORLOWSKI, MD, MACP, Chief Health Care Officer, Association of American Medical Colleges
BETTY RAMBUR, PhD, RN, Routhier Endowed Chair for Practice and Professor of Nursing, University of Rhode Island*
JOHN RATLIFF, MD, FACS, FAANS, Associate Professor of Neurosurgery; Vice Chair, Operations and Business Development; Co-Director, Division of Spine and Peripheral Nerve Surgery, Department of Neurosurgery, Stanford University Medical Center, American Association of Neurological Surgeons*
SRINIVAS SRIDHARA, PhD, MHS, Managing Director, The Advisory Board Company
BILL WEINTRAUB, MD, FACC, Chair of Cardiology, Christiana Care Health System*
HERBERT WONG, PhD, Senior Economist, Agency for Healthcare Research and Quality
DOLORES YANAGIHARA, MPH, Vice President, Performance Measurement, Integrated Healthcare Association

NQF STAFF:
TAROON AMIN, MPH, PhD, Consultant
HELEN BURSTIN, MD, MPH, FACP, Chief Scientific Officer
HIRAL DUDHWALA, RN, MSN/MPH, Project Manager
ERIN O'ROURKE, Senior Director
RACHEL ROILAND, MS, PhD, Senior Project Manager
IRVIN SINGH, MPH, Project Analyst
ASHLIE WILBON, RN, MPH, Senior Director
MARCIA WILSON, PhD, MBA, Senior Vice President, Quality Measurement

ALSO PRESENT:
CHAD HEIM, Vice President, Health Informatics, HealthPartners*
GARY KITCHING, Senior Manager, Health Informatics, HealthPartners*
SUE KNUDSON, MA, Senior Vice President, Health Informatics, HealthPartners*
SRINIKETH NAGAVARAPU, PhD, Senior Policy Associate, Acumen
KIMBERLY SPALDING BUSH, OT, Director, Performance-Based Payment Policy Group, Center for Medicare, Centers for Medicare & Medicaid Services
SCOTT WETZEL, Senior Specialist, Quality Reporting, Association of American Medical Colleges

*** present by teleconference**

C O N T E N T S

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

(9:10 a.m.)

DR. ROILAND: All right, good morning everyone and welcome to today's meeting of the Cost and Resource Use Standing Committee. My name is Rachel Roiland, and I'm the Senior Project Manager for this project and, along with the rest of the staff up here today, we'll be guiding you through this evaluation of three measures today.

But before we do that, I just want to provide some housecleaning details. Irvin, if you could go to the next slide please? All right. We'll just wait until it loads there -- all right.

Just so you all are aware -- if you could go back one slide Irvin -- there are restrooms available on this floor of the building. If you go outside this room and go straight past the elevators and hang a right, that's where you'll find the restrooms.

We do have two breaks scheduled for today, the first at 10:45 in between our

evaluation of the two health partners measures, and then lunch is also provided today by NQF.

For those of you with laptops and cell phones in the room, we do have a WiFi network available for you and the log-in and username and password are available upon the screen.

We do ask those of you who have cell phones if you could please mute then during the meeting just to minimize distractions.

And if you do need to take a phone call you're more than welcome to do that, we just ask that you go out into the area where you picked up your name tags, just to minimize the background noise when we're having discussion.

For those of you in the room, we do ask that you use your microphones when you're talking and talk directly into the microphone, just so that it helps the folks who are listening in to better hear everything going on.

And it also helps our court reporter take a better transcript. And also just a little note -- there can only be three microphones on at

one time, so when you're done speaking, please make sure to turn your microphone off.

And then when you do want to speak, we ask that those of you in the room put your name tents up, like this. It just helps Brent and Cheryl know who to call on.

And then for those folks on the phone, or participating in the webinar, we ask that you raise your hand by the webinar platform, and that will signal to us that we need to let Brent or Cheryl know that you'd like to jump in and be a part of that conversation.

So that's all the housekeeping stuff, and now I'll turn it over to Erin O'Rourke, our Senior Director, to just give us some opening remarks, and we'll introduce the rest of the staff.

MS. O'ROURKE: Thank you so much Rachel, and thank you to all of you for braving the inclement weather and getting down here, or for those of you who are on the phone, for bearing with us for an extended web meeting. We

really appreciate everyone's efforts to come together today and do this work, so thank you.

So I think with that, I can just introduce myself. I'm Erin O'Rourke, the Senior Director supporting this project and I'm excited to work with you all, joined by obviously Rachel, our Senior Project Manager, Hiral Dudhwala, our Project Manager, and Irvin Singh, our Project Analyst. We also have Taroon Amin, consultant for NQF helping to support this work. I think many of you have worked with him in the past. We're joined also by Helen Burstin, our Chief Scientific Officer, and Marcia Wilson, our Senior Vice President. So that's the NQF team, and I think with that I can turn it over to Brent and Cheryl to say a few words.

CO-CHAIR DAMBERG: Great, thanks everyone for joining us in person today, and for those of you on the phone, I know it's going to be a long day but thanks for joining us and hanging in there throughout the day.

We have three measures that we're

going to review today and I know all of you spent a fair amount of time looking through them and probably have a lot of interesting comments to provide, and so we'll look forward to that discussion.

CO-CHAIR ASPLIN: Good morning everyone. Brent Asplin. It's good to see a lot of familiar faces in the room, and look forward to the conversation today.

And background, emergency physician, I was recently at Mercy Health in Ohio and Kentucky as the Chief Medical Officer, and look forward to the conversation and the dialogue around these methods this morning and this afternoon.

And with that I'll turn it over to Marcia Wilson, and we can go through the declarations of conflicts.

DR. WILSON: Thank you so much, Brent. As Erin said, I'm Marcia Wilson, Senior Vice President for Quality Measurement, and I'm filling in for our general counsel today.

It's our custom here at NQF to combine introductions and disclosures of interest. And I know when you signed up for this committee, you got a very long DOI form, disclosure of interest form.

And today we're going to ask you to orally disclose any relevant activities, relevant in terms of the work before this committee.

So, it's not necessary to summarize your resume, you are a stellar committee, and so we bring considerable expertise to this panel, but we do want to know if you have any activities, funded or unfunded, that are directly related to the measures and the issues coming before the committee today.

So, just a couple of reminders. You sit on this committee as an individual, you don't represent your organization. So, for example, when you do your introduction you would say, hi I'm Marcia Wilson and I work at the National Quality Forum.

So you are seated as an individual,

and just because you disclose something, it does not mean that you have a conflict. But we do this is the spirit of transparency and openness. So, what we are going to do is start with introductions and disclosures.

Here in the room, I'm going to start with our Co-Chairs. We'll go around the room and then I know we have quite a few committee members on the phone today, and I will call your name and you can do an introduction and disclosure once we're finished in the room. So Cheryl, if you won't mind starting please?

CO-CHAIR DAMBERG Sure, Cheryl Damberg, I'm the Senior Researcher at the RAND Corporation. I have a number of contracts with CMS. One of them in particular is around the Medicare star ratings, which involves us analyzing performance measure data that's submitted by health plans to Medicare and generating the star ratings. But other than that I don't think I have any conflicts related to these three measures.

CO-CHAIR ASPLIN: Brent Asplin, currently independent. I do not have any disclosures, although I do think the committee should know that I worked as an Emergency Physician for Health Partners Medical Group about a decade ago, but did not have any direct involvement in the development of the measures we're going to be reviewing today.

DR. WILSON: Thank you Brent. And if we go to Janis, next.

MEMBER ORLOWSKI: Good morning. I'm Janis Orlowski, and I am the Chief Health Care Officer working at the AAMC, and I have nothing to disclose.

DR. WILSON: Thank you.

MEMBER NAESSENS: I'm Jim Naessens, I'm a Health Care Researcher at the Mayo Clinic and I have nothing to disclose.

DR. WILSON: Thank you.

MEMBER SRIDHARA: I'm Srinivas Sridhara I work for The Advisory Board, and I have worked with all three measures actually,

part of grant programs in the past and both implementing and testing for the health partners measures, along with Henry so that, though I'm no longer engaged in that and have no role in the measure development process.

And for the Medicare spending for beneficiary, we've tested and implemented in both the State of Maryland, and now The Advisory Board. So, but again, no association with measure developers currently.

DR. WILSON: Okay.

MEMBER NEEDLEMAN: Morning. I'm Jack Needleman, and I'm a Professor in the Department of Health Policy and Management at the UCLA Fielding School of Public Health.

I don't think I have any conflicts but in the interest of full disclosure, as Marcia said, I'm sitting as a member of the technical expert panel that CMS is using to evaluate the trends in their quality measures over time, and I believe this measure today is included in the purview of that committee.

DR. WILSON: Thank you.

MEMBER GARRETT: Good morning, I'm Nancy Garrett, I'm the Chief Analytics Officer at Hennepin County Medical Center, which is a safety net provider in Minneapolis.

And I don't have any conflicts to disclose other than just being aware that the HealthPartners measure is used in Minnesota through our Minnesota community measurement organization, and I've been involved in kind of that roll out and using it as a provider as well.

DR. WILSON: And again, using measures does not present a conflict.

MEMBER WONG: Good Morning. My name's Herb Wong. I am an economist with the Agency for Healthcare Research and Quality and I have nothing to disclose.

MEMBER MARCINIAK: Hello, I'm Martin Marciniak, I'm a Vice President with GlaxoSmithKline and a policy economist by training. I worked in the U.S. medical affairs organization.

MEMBER YANAGIHARA: Good Morning I'm Delores Yanagihara, I'm Vice President of Analytics and Performance Information at Integrated Healthcare Association in California, and I have nothing to disclose.

MEMBER EAMES HUFF: Hi, my name is Jennifer Eames Huff, I'm an independent consultant and participating in this on behalf of the Pacific Business Group on Health. I have nothing to disclose.

DR. WILSON: Thank you, and we'll now go to the committee members on the phone, alphabetical just to give you a heads up. We're starting with Larry Becker, are you on the phone?

MEMBER BECKER: I'm here Marcia, I'm Larry Becker. Recently retired from Xerox. I'm on the Board of NQF and I have nothing else to disclose.

DR. WILSON: Thank you. Mary Ann Clark?

MEMBER CLARK: Yes, good morning. It's Mary Ann Clark. I'm a Vice President of Health

Economics and Advanced Analytics at Avalere Health. Although I personally have nothing to disclose, I believe in the past some others in our company have worked on behalf of our client to develop quality measures for submission to NQF, but I have not worked directly on any of those and that was in the past.

DR. WILSON: Thank you. Is it Troy Fiesinger?

MEMBER FIESINGER: Yes, this is Troy Fiesinger. I'm a physician in balmy Houston, I have no conflicts to disclose for today's meeting.

DR. WILSON: Thank you. Andrea Gelzer?

MEMBER GELZER: Hi, Andrea Gelzer, and I'm Corporate Chief Medical Officer at AmeriHealth Caritas, and I have nothing to disclose.

DR. WILSON: Thank you. Kristine Martin Anderson?

MEMBER MARTIN ANDERSON: Hi, I'm

Kristine Martin Anderson, I'm with Booz Allen Hamilton. I have no conflicts, but Booz Allen does have a number of contracts with CMS, OCM, VA and others, where we are selecting measures for use in new programs, which include efficiency measures.

DR. WILSON: Okay, thank you. Betty Rambur?

MEMBER RAMBUR: Hi. Betty Rambur, Routhier Chair for Practice and Professor of Nursing at University of Rhode Island and very recently moved from Vermont where I was a member of the Green Mountain Care Board, and I have no conflicts of interest to disclose.

DR. WILSON: Thank you. John Ratliff?

MEMBER RATLIFF: Hi, good morning. I'm John Ratliff, I'm a neurosurgeon at Stanford. I've done some consulting and collaborative work with Acumen one of the developers on episode groupers for CMS, but I have no conflicts relevant to these measures.

DR. WILSON: Thank you. Bill

Weintraub? Bill, are you on the phone? Do you show him signed up?

Okay, Bill, it may be that you're signed onto the webinar but not called in, so I'm going to go ahead and keep moving on with the DOIs, and then if you are able to join us by phone then you give a -- introduce yourself and give us a disclosure.

And I think -- is there anyone else on the phone that I have not called? Okay, just a final comment that if at any time during the meeting you think that you have a conflict of interest, or if you perceive that someone else had a conflict of interest, we do ask that you speak up.

You can approach our Co-Chairs, or any of the NQF staff. What we don't want is for you to sit there and think that there's some conflict going on and not bring it to our attention in real time as opposed to waiting later.

So, based on what you've heard from your colleagues on the committee, and from the

remarks, does anyone have any questions? Okay, thank you very much.

DR. ROILAND: All right. Thank you Marcia. Hi everyone, again. This is Rachel for those of you on the phone, and I'm just going to give you a brief introduction to the project that we have today, as well as give a little bit of historical context of where we've been and how we've gotten to where we are today.

This will be old news to all of you who worked on most of these projects but just to, sort of, touch base again on them.

NQF's prior work in cost and resource use has really focused a lot on doing a lot of conceptual work and trying to come up with frameworks that we can utilize to evaluate these measures, as well as in applying those evaluation criteria that were developed to various cost and resource use measures.

So here up on the slide we just list those various projects that a lot of you are most likely involved in.

And so we begin with Episodes of Care Measurement Framework, which really laid the groundwork for how we look at episodes of care and apply quality measurement to those episodes, and then --

MEMBER WEINTRAUB: Excuse me. Bill Weintraub, I was only on the webinar, my headset didn't work, but I'm now on the phone.

DR. ROILAND: Oh, thank you Bill. And did you have anything to disclose?

MEMBER WEINTRAUB: Yes, well I do have -- I'm a cardiologist, Christiana Care in Delaware, outcomes researcher. I had a CMMI outcome support --

(Telephonic interference.)

MEMBER WEINTRAUB: -- included.

DR. ROILAND: Okay, thank you so much Bill. We appreciate it. All right, so just to bring us back to talking mode and NQF's prior work in this area.

So there's been three phases before this phase where we've applied the cost and

resource use evaluation criteria to various measures, and this fourth phase is actually our first maintenance phase.

So we'll be taking three of those measures that you all have reviewed before and putting them through the maintenance process for NQF. So it's a bit of a new territory, a mix of old and new. The measures you've seen before, but the maintenance process is new for these particular measures.

And then there's been various other projects talking about different methodological challenges when developing and evaluating cost and resource use measures as well.

So, you all have done a lot of work thus far, and will do a lot of work today as well. So, we look forward to it, and thank you all again for being here today.

Go to the next slide. This slide just lists our current portfolio of NQF endorsed measures for cost and resource use measures.

And this also includes the three we

will be looking at today which include the first two, the Total Resource Use Population-based index, as well as the Total Cost of Care index, and the one listed on the bottom, the Medicare Spending Per Beneficiary hospital.

Go to the next slide. And this just reiterates the same point that this project today will be evaluating measures related to cost and resource use that can be used for accountability in public reporting for the population that they've been specifically tested for.

And so, again, the three measures we'll be reviewing today are listed at the bottom of the slide there. So, three measures today, but we expect it to be a rich discussion, so we look forward to hearing all of your thoughts.

And with that, I will turn it over to our Project Manager, Hiral, who will walk us through the evaluation process we'll be going through today, as well as voting procedures. So Hiral, over to you.

MS. DUDHWALA: Okay, thank you. So

again, this will probably just be a refresher for all of you. Next slide we're just going to go over a quick overview.

So, the roles of the standing committee during our evaluation meeting today is to act as a proxy for the NQF multi-stakeholder membership to work with the NQF staff to achieve the goals of the project.

Evaluate each measure against each criterion. Indicate to the extent to which each criteria is met and rationale for the rating. Make some recommendations regarding endorsement to the NQF membership.

Oversee the portfolio of our cost and resource use measures, as well as select two to three year terms for our new committee members.

Next slide, please. All right, and also just some ground rules for our meeting today.

So during our discussion, committee members should be prepared, having reviewed the measures beforehand, base evaluations and

recommendations on the measure valuation criteria and guidance.

Remain engaged in the discussion without distractions. Attend the meeting at all times except breaks. Please keep comments precise and focused. Avoid dominating a discussion and allow others to contribute. Indicate agreement without repeating what had already been said.

For those that are participating on the phone, please make sure to use your raise your hand function and we will be monitoring and call out your name so that you may also participate.

Next slide. Okay, so again review the measure evaluation process. You have received a measure worksheet which includes a preliminary analysis.

The committee pre-evaluation comments are included in that, as well as the pre-meeting public and member comments, as well as any evidence and testing attachments.

We just want to stress that -- NQF's

staff would like to stress that preliminary analysis is intended to be used as a guide to facilitate the committees discussion and evaluation. Next slide please.

Okay, so measure discussion and voting. There will be a brief introduction by the developer, about three to five minutes to begin with. There are lead discussions which will begin the committee discussion by providing a summary of the pre-meeting evaluation comments as well as emphasizing areas of concern or differences of opinion.

The developers will also be available, whether in person or by phone, to respond to any questions. The full committee will discuss and then vote on the criterion before moving on to the next criterion.

So voting process for today. The voting tool. All in person voting members have a remote clicker to vote. You should all have it right next to you.

For those that are on the line, we

have already communicated with you how you can send your votes to an NQF staff member and they will do the clicking for you here. All voting members, as I just said, who are remote, we have those instructions already provided to them.

So, with the clickers, make sure to point the clicker towards our staff member here in the corner, Irvin. When voting, the remote will briefly display your vote choice.

You do have the chance to change your response, so whatever is your last option that you choose on your clicker is the voting that will be the final vote. So, you can change it if you do make an error, so --

Okay, next slide. All right, so, for achieving consensus, quorum would be 66% of the committee, so that would be, for this meeting, 13 committee members.

Pass -- recommend greater than 60% of the committee who vote yes, which would be High plus Moderate. Consensus not reached would be 40-60% yes votes, inclusive of 40 and 60%. So does

not pass, not recommended, less than 40% yes. Next slide.

Okay we just went over quite a few items, so if there any questions from anyone in the room or by committee members on the phone?

Okay. All right. Next slide. Okay. Well, I will go ahead and pass it on to our Co-Chairs, as we begin the candidate measures. Brent and Cheryl.

CO-CHAIR ASPLIN: Very good, thank you, Hiral. And good morning, we're going to start with Measure 1598, the Total Resource Use Population-Based Per Member Per Month Index. The developers, Health Partners.

And we'd like to begin by inviting the developers to provide a brief overview of the measure, and I'd ask them to introduce themselves. I'm not sure if the developer is here or on the phone.

DR. ROILAND: They're on the phone, Brent. So if, Operator, if you could open their lines, if they are not already open.

MS. KNUDSON: Hi this is Sue Knudson from HealthPartners, can hear me okay?

CO-CHAIR ASPLIN: Yes, good morning, Sue. We can hear you just fine.

MS. KNUDSON: Hi, okay thank you. Just to let you know, I'm joined by my colleagues, Chad Heim, Gary Kitching, Erica Vetta, and Kim Ritten.

So I'm going to do just a brief introduction, and before I do that I just want to say thank you to the NQF staff and to all the committee members. We know this is a big time commitment for all of you, and appreciate the attention. Both the NQF as well as all of you individually are going to address this important topic in health care.

So first, about HealthPartners. We are a consumer-governed non-profit organization. We're both an integrated health care and financing organization with about 2300 team members located mostly here in Minnesota.

Our health plan has about 1.5 million

members. We're ranked by NCQA among the top 30 in the nation. From a care group perspective, we care for about a million patients. We have about 1700 physicians that cover multiple specialties.

Six hospitals including a level one trauma center in the heart of St. Paul, all the way down to community critical access hospitals in size.

Our mission is to improve the health and well-being of those we serve in partnership with our members, patients and the community. And our vision is an important one, because it helps you to understand where this work comes from, health as it could be, affordability as it must be through relationships built on trust.

So we're also an organization, and a group of team members, that is committed to health equity and serving community needs, so to that end, that includes providing and advocating for mental health services.

And we are number two in the state for charity care services, and number one in the East

Metro of the Twin Cities.

So, just quickly about our measures. There's two measures that we'll be reviewing today: total cost of care and resource use, starting with resource use. They are for a commercial insured population. They're based on all care that's provided to patients, so we refer to them as a population based measure approach.

We developed them upwards to 20 years ago on the TCOC side, and in the early 2000's created the algorithm that creates the resource use measure that we will be covering first. The costing technique is the only difference between the two.

So for total cost of care, it's the allowed payments, which is inclusive of what the plan pays as well as what the member pays.

But for resource use, where we're starting, that uses our standardized patented algorithm for standardizing the cost approach. So, we'll talk I'm sure a bit more about that.

The measures are powerful when used

together, because by having the two you can do the simple math to understand price drivers as well. However, resource use on its own is also very powerful in helping to drive care delivery improvement by steering, in a focused way, where there are pockets to improve in the overall performance.

So again, their commercial members are measured. We have as an organization calibrated for Minnesota, for our own use, a Medicaid model, but that is calibrated, we've been under a bidding system in the state for several years now that is not translatable to other states, therefore we did not submit it.

We are a Medicare cost plan, planning to transition to Medicare Advantage, so we do not have the full suite of Medicare claims, so therefore we could not calibrate at this time a Medicare model.

And if users have those data, we still recommend segmenting the measures among the different populations to account for differences

in disease prevalence, enrollment profiles and the like.

So what I think is important in terms of context for the committee to understand is we do measure everything from a particular point of view here at HealthPartners, so our quality results from a care group point of view lead the state at community measurement, and we've been able to make improvements using these measures.

At the same time, our quality has improved, we're assisting at very high levels, and for a health plan point of view, the same goes with being ranked in the top 30 in the country.

So since the initial endorsement in 2012 before we get into the measures, just to close my comments. The spread and the uptake of the measures has been impressive. We have well over 200 users across the country. Those types of users -- manned health plans, providers, consultants, and measurement collaboratives, as well as researchers.

We've enhanced our public domain website beyond just containing the specifications. We also have out in the public domain the software to actually create the measures. We've also included software that, if a user would like, they can create the full suite of reports that HealthPartners produces for our network providers as well.

And one thing I forgot to mention about. This is just not a health care or a HealthPartners set of measures. We do the same reporting for every contracted partner in our health plan network, and work with them, because our attitude is their success is our success.

So in terms of usability, there's been a strong uptake, and we've gotten very favorable results by being very transparent and putting all of our tools, if you will, out there because we're an organization that wants to compete on our results, not just the techniques.

So several of those users include, and you guys probably saw the comments from several

of the delivery systems that we work with. For example, Sanford, or Ascension Health, or Fairview. And then also, the national Network for Regional Healthcare Improvement and all of their sites in different states cross the country.

And Nancy mentioned in her comment, the uptake here at Minnesota community measurement. So the work has been really great to be a part of, and we're proud of this spread, and it's what we can do as a consumer-governed non-profit to contribute to the issue of affordability in healthcare. So, that would conclude my comments, Brent.

CO-CHAIR ASPLIN: Very good, thank you, Sue. I appreciate your comments and being on the phone here this morning. And we're going to move forward then with the review of the measure, and I'll just -- for review, for many of you, but we have some new committee members.

Just want to -- for efficiency's sake, stay disciplined about the focus of the

conversation today. Because for each measure, if it doesn't pass on importance to measure or scientific acceptability -- those are must pass criteria -- so we want to do it in order and restrict our comments to the topic.

We will vote after each section. There will be plenty of time at the end of the conversation, if we get through all the areas, for summary comments from committee members, if you wish, before our final vote, which is a recommendation to endorse or not to endorse after we've gone through all the different sub categories.

So with those comments, let's get started. On 1598, Larry Becker is our lead discussant for importance to measures, so Larry, the floor is yours.

MEMBER BECKER: Thank you very much. Can you hear me?

CO-CHAIR ASPLIN: Yes, we can hear you just fine. Thank you.

MEMBER BECKER: Good. Thank you.

And so one thing I would note in the document, that the developer, to marry the two measures together on importance. But let's begin with 1598. I think everybody would agree that cost and affordability is a major concern in the healthcare system. And it contributes to the number of uninsured. It contributed to budget deficits. It contributes to medical bankruptcy.

Before you can understand how to effectively lower costs without decreasing quality you need to understand the total cost of virtually everything that goes into the care equations here. And one of the things that I noted in the developer's discussion of this is that most, if not all, of those contributors are inside these indices around primary care: so, professional; facility; in-patient and out-patient; pharmacy; radiology; lab; ancillary and behavioral health. There may be more but that seems to me that those are the biggest contributors to these things.

And in my experience in industry,

having worked for Xerox for nearly 30 years, is that one of the ways you begin to think about cost is you look at something called UMCs, unit manufacturing costs. In other words, what are all the contributors to developing an output. You know, in this case it might be an office visit. It might be a surgery. It might -- whatever it is.

And that's the only way in industry that we are able then, once we understand the big costs, is to look at all the contributing elements, all of the inputs to that and begin to break it down, and look at the individual pieces to understand where inefficiencies are, where there are opportunities, as opposed to, you know, just sort of saying arbitrarily we're going to change this or that.

The other really important thing here that we learned in industry is something called entitlement. It's a Lean Six Sigma term. It's probably not what you're thinking. It's not that people are entitled to A or B. Entitlement in

the Lean Six Sigma context says that you are able to compare what you do to what else is being done by similarly situated people or processes out there, so that if somebody is producing that output at, say, 100 and you're at 90, you know that you could, given the right input, be able to get to that same hundred. Somebody else has achieved it. And that's what they mean by "entitlement."

So, by being able to look at total costs and looking at people, benchmarking across multiple systems, multiple providers of the same type, you begin to begin to be able to make comparisons and say, I'm here; it's possible for me to get there.

And so, the importance in this measure is to begin the conversation and to begin to sort of, if you will, level the playing field so that people can make appropriate comparisons to be able to take appropriate action.

So, I'm going to stop there and ask if others wanted to add to that.

CO-CHAIR ASPLIN: Thank you, Larry.

Any comments from committee members on this area? We'll have three votes under the overall importance. The first will be on the priority, our assessment of the priority of the measure.

The second would be on opportunity for closure in care gaps.

And then the third is around the measure intent.

So, are there any comments from committee member at this time? Martin.

MEMBER MARCINIAK: So I'm echoing Larry's concerns. I mean, one of the things that I saw when I reviewed this measure, and reviewing it I guess now three years ago, was a general concern that when things are aggregated it speaks to Larry's point: how do you disaggregate it to really find quality improvement?

That was something I was looking for when I read through the review this past time. Didn't really feel I saw that. It would be nice

if at some point the measure developer might comment on that.

CO-CHAIR ASPLIN: Sue, do you have any comments in response to that?

MS. KNUDSON: Yes. Sure, I'd love to. Thank you for those questions and comments.

What we have found to be very useful -- and you could see this on our public website as well as in the Minnesota Community Measurement work, for example, they've taken a different approach. But there is drillability to the measures. And so what we've provided in our reporting suite is a dashboard, if you will, that include this measure and then companion utilization measures that go along with it.

And so, for example, as an organization on our own, as we are working with providers we will benchmark back to our plan average performance. And not that folks aspire to be average, but they can set goals. It makes the map easy and they can set goals that might be 10 points better than average, for example.

And then by pointing to those risk-adjusted utilization measures that would flow right in the same reporting suite -- and by the way, this is the software also that we've put out on our public domain website for others to replicate -- the measures just begin to cascade down. And so you can understand do you have, compared to your peers, an opportunity with admissions, do you have an opportunity with high-tech imaging used outside of the ER at higher rates; so, those major contributing factors that drive the overall performance. So we've made that available.

Does that help?

MEMBER MARCINIAK: It does for me. Thank you.

CO-CHAIR ASPLIN: So, go to Jack next. And then, Troy, I see that you've raised your hand. You'll be following Jack.

MEMBER NEEDLEMAN: Okay. It probably reflects the fact I've been doing this too long. But somewhere in the documentation is a

reference, and it may -- so this is a quality measure. And so I just think because it always permeates our discussions, comes up at some point or another, I think it's important to note at the beginning this is not a quality measure. Cost and resource use are not measures of quality. Nor is it an efficiency measure, which is where people ultimately want to go is Are you more or less efficient?

But, as we have discussed in the past, more or less efficient is contingent upon how your resource use translates to your performance on other kinds of measures of quality or outcomes. And without that mapping and without that examination of the correlation, it's hard to talk about efficiency here either. So, this is an input to the discussion about efficiency as the developer, I think, pointed out.

And I think one of the issues of importance is, by itself the resource use measure has some interest but is not getting all the way we want to go. So, the question I would ask the

developer is the extent to which they have provided mapping tools for looking concurrently at their outcome measures and quality measures, which are also included in the suite of things they do, and the resource use or cost use measures that make it up.

CO-CHAIR ASPLIN: Thank you. Sue.

MS. KNUDSON: Chair, would you like me to address that?

CO-CHAIR ASPLIN: Yes, please.

MS. KNUDSON: Yes. So, we have take -- we totally agree with what you are saying, this is one of the building blocks that would help drive to that overall efficiency or value equation. We have taken a multi-step approach. And so when we display information we generally do not display cost information on its own, it would always be paired with quality and experience information, if you will.

But we look at those from two different bars. So, for example, in our shared savings contracts with our provider network, you

must meet the thresholds that we have in place for quality and experience assessments. And then if you perform well in this cost of care and efficiency, then you would, you know, benefit by having bonus and/or shared savings payments paid out.

We use the same approach with transparency on our website. We offer both a quality and experience score for consumers to use, as well as that being paired with the overall cost information. Because the literature is pretty solid that there is no direct correlation between cost and quality, we have actually not take it that step further and done a value equation, which I think is perhaps the spirit of where you're heading in the discussion.

And we have done that because sometimes the map doesn't quite work out. And consumers, particularly, need to make choices based on where they're at in a particular point of care. And sometimes that quality information on its own might be the driving force and/or the

costing combination.

So that is the way we have used it.

CO-CHAIR ASPLIN: Thank you, Sue. Troy and then Andrea.

MEMBER FIESINGER: Yes. This is Troy Fiesinger from Houston.

You already addressed part of my question about drillability. But in my mind I want to make sure I understand clearly. How low can you go? I mean, if I have a 4- to 10-doctor group over a 50-doctor practice and then 500 physicians in an ACO, how long in unit can you get down to get to applicability and usability?

MS. KNUDSON: So, we're asking for endorsement on the measures at 600. And the reason we're asking at 600 even though we've tested and they're, they're good at lower levels, is because that's the sweet spot where we have found we have that robust quality and experience information available for all those uses that I had mentioned.

They are -- we have done testing --

they are reliable and valid at lower levels. We also, in terms of the drillability, to your point in the question, we can drill that all the way down to the physician level. But what I will tell you is that from, you know, a real lift in improvement point of view, when we consult and work with groups we have found the best way to drive improvement is recognizing the fact that all of us as individuals, care team members included, come to work wanting to do the best they can. And we have often and predominantly found that putting in place system solutions, whether they be computer solutions or process solutions across the team, are the best way to list performance and drive results that you can see.

So, we really take that approach with it and have found very few individual physician opportunities. And that's largely recognizing, yes, we are largely a group practice state. But we do have smaller groups and single-specialty practices that we work with as well. But

generally it's the same types of solutions.

So, you can let me know if you still have outstanding questions based on that. I'd be happy to answer them.

MEMBER FIESINGER: Thank you. That answered the bulk of my questions. Also to be able to aggregate your data by payer contract.

MS. KNUDSON: Yes, we can. If I put my care group hat on, you know, using the Minnesota Community Measurement, that is all aggregated together, you know, from a usability point of view. Our care group colleagues would get reports from each of their payers which would directionally show different levels of performance.

But these measures themselves, they are drillable down to conditions, procedures, et cetera. So you can also understand those drivers. And what we have heard from care groups is directionally, you know, the different payers may be kind of honing in on the same things using a slightly different methodology. But that's the

credibility about these measures is that when we bundled all of our data together at Minnesota Community Measurement it's pretty much singing the same song. So the standardization of the approach has been really useful.

CO-CHAIR ASPLIN: Thank you. Andrea and then Jennifer.

MEMBER GELZER: Hi. Thank you.

I would just note that this is a widely-adopted measure, even by state Medicaid. We do it in multiple states. And we're seeing it increasingly being adopted and, you know, have supported it.

And it's a given that this wouldn't be used in a vacuum without a, without quality, you know, metrics. And the quality -- and to me, as a resource use measure, you know, it's the start. It's an overall global cost and then you -- or and then you need to go digging. And you have to have the data, obviously, to do the digging. But more and more we are able to do that, and not just payers able to do that, but providers as

well.

The quality metrics are driven by -- that we would pair with this measure are, you know, driven by the population and driven by the issues that we observe in the population. I would say that, you know, let experience with the measure inform it and improve it going forward. But it's a valid measure.

And I would ask the measure developer how -- what are the main categories and areas that were changed from the beginning to the current iteration, if any?

MS. KNUDSON: Yes. The only difference is really in our truncation limit. We boosted that from 100,000, or our Winsorizing, to 125. And that is just in recognition of inflation and where medical costs are at. So that is really the only substantial change.

MEMBER GELZER: Thank you.

CO-CHAIR ASPLIN: Thank you. Jennifer.

And then one quick comment. For those

of you on the phone using the raise-hand function, after you've given a comment or asked a question if you could lower your hand so we know whether you're asking to speak again or not. Thank you.

MEMBER EAMES HUFF: Hi. I just want to say I agree with all the comments so far that have been talking about the support to the team on a high priority measure. I want to drill down a little bit more on the performance gap area. And I think the evidence that we showed does indicate there's variability.

My question is more about since this is a relative score, how does looking at the, for example, the medical variability that's included here, how does looking at medical group variability from year to year, how does that adjust since it is a relative score in terms of thinking about how improving opportunities are having an effect?

MS. KNUDSON: Yes, I'll start and then ask one of my colleagues, if I don't cover it

sufficiently, to add on.

We do that in a couple of ways. We always index performance to the current year. And in order to -- so we understand where any group or drill level is performing relative to current performance of peers.

However, we also in our reporting suite offer the capability to index those last two years previously also to the current year, so you can always see how you're moving over time. And all of that's available in a class of different measures as well.

Does that help, Jennifer?

MEMBER EAMES HUFF: Yes. Thank you.

CO-CHAIR ASPLIN: Let's see, Jack, probably the last time.

And then we're going to move to our first vote in this area.

MEMBER NEEDLEMAN: Okay. Well, this is interesting because Jennifer read the variance in these measures differently than I did. And what I'm about to say, you know, I think it's

important, let me start out by saying I think it's important to have a measure of resources, measured resources over time given the goal for efficiency.

But I found that when I look at the variance on the resources measure, not the cost measure but the resources measure, the standardized pricing model, with the exception of a few groups that are at the tail that seem to be more extreme, folks seem to be pretty tightly clustered. And when I look at the report you provided of the 3-year data on the groups, it looks like about half the groups are going up over three years and half the groups are going down, none by very much on this measure.

So, since one of the elements of importance is room for improvement, and variation in performance, I'm wondering if you can provide a little bit more insight into how much variance are you actually seeing in this measure, leaving aside the three or four or five groups that are at the extreme tails, which could be very

peculiar patient experiences that year for those groups?

But how much variance are you seeing and how much improvement are you seeing over time, not in the index but in the actual measure itself, which is a measure of resource use? How much change or trend over time are you seeing? And how consistent is that across the groups?

MS. KNUDSON: Yes. Thanks, Jack, for that astute question. Because we're not just managing for the tails. If we did that, we wouldn't be making an improvement. And really to sort of be one with the results of these data is really important for the different users and these folks who are being measured. Because I will give you an example from our own delivery system, our west side delivery system here in the Twin Cities.

Even a point of improvement in resource use is about a $4 million improvement in affordability that we can return to patients and purchasers. So even the small, even around the

averages, it's really important to really drill into the data and look for performance. It is very hard to get like a 10 point improvement.

So that is not how the numbers should be interpreted. And so we're really looking to those utilization metrics and other things that help with usability to drive those. It's sort of a steady pace of an improvement portfolio. And even at a half a point improvement, that is a lot of dollars to be returned.

CO-CHAIR ASPLIN: Thank you, Sue. Janis, a quick comment?

MEMBER ORLOWSKI: I want to raise concerns regarding the disparities adjustment in this measure. And --

CO-CHAIR ASPLIN: Yes, can we table that? Because that I know is going to come up later on. But if could just hold off and get into our votes on importance, if you don't mind. And remind me, so I don't, don't miss --

MEMBER ORLOWSKI: But you don't think it's critical to the importance? You feel it's a

separate subject?

CO-CHAIR ASPLIN: I think it's going to be in the validity section. So it should be tied to the second vote.

All right, very good. With that, let's move the first vote in importance to measure is under the category priority.

And, Rachel, how are the folks on the phone going to vote? Do you want to quick go over how we're going to handle that today?

DR. ROILAND: Sure. So, hello, everyone. So for the folks who are participating on the phone, we have reached to you individually and told you who of the staff is going to be your proxy. And they will be emailing or texting us, whatever works best for them. So that's the process that will be happening.

And we have our clickers labeled so we know who's getting what.

CO-CHAIR ASPLIN: All right. So, good. And any questions from those on the phone? We are ready to roll.

MEMBER WEINTRAUB: Yes, I have a question. Can you hear me?

CO-CHAIR ASPLIN: Yes, we can.

MEMBER WEINTRAUB: So if I was to look and get an email from someone as my proxy, I don't see it.

CO-CHAIR ASPLIN: Is that Troy?

MEMBER WEINTRAUB: It's Bill Weintraub.

CO-CHAIR ASPLIN: Bill. I'm sorry, Bill.

MEMBER GELZER: And Andrea Gelzer. I'm having a hard time locating the specific email.

DR. ROILAND: Okay. We'll send an email right now. Just one moment, please.

(Pause.)

DR. ROILAND: Bill and Andrea, if it works better for you, you can actually just chat it to us in the webinar platform. Would that work for you?

MEMBER GELZER: That would be great.

MEMBER WEINTRAUB: Yes, well let's try it and make sure it works. We'll try on this first go-around.

MEMBER FIESINGER: This is Troy. I think chatting it would be easier for me, too. Then I don't have to jump between screens on the computer.

DR. ROILAND: Okay.

Irvin, can you set up the voting slide, please.

MR. SINGH: Yes, absolutely. So --

MEMBER WEINTRAUB: All right, so you want us on the phone to put it in right now?

DR. ROILAND: Let Irvin read out what we are going to be voting on and the voting options. And then once he is done with that he will ask you to submit your votes. And folks can, who are on the phone if you prefer you can submit it through the webchat. Or if you want, you can submit your vote through the email or text that we had arranged before. But whatever works best for you, we'll figure it out up here.

MEMBER WEINTRAUB: All right. So, just give us feedback that it works after this first vote; okay?

DR. ROILAND: Sure. We definitely will.

MR. SINGH: Good morning, everybody. We are going to commence the voting process. And just want to make sure that the voting platform is working okay, making sure that your votes are being captured.

The first thing we're going to do is run a test vote to make sure that the clickers are working and that you're able to submit a response. So go ahead and submit your test vote now. And just point your clickers to this direction because this is where the receiver is.

DR. ROILAND: And for folks on the phone, you can't see the voting slides. We're not able to show them on the webinar platform. So the test slide, your voting options are 1 for high; 2 for moderate; 3 for low; and 4 is for insufficient.

So, if you could submit those via the method that works best for you we'd appreciate it. Thanks.

MEMBER WEINTRAUB: All right. So it's not clear to me how to use this chat function. I typed "chat" and what I see here is several messages. How do I respond back with a chat?

MEMBER BECKER: At the bottom -- this is Larry -- at the bottom you see a box that says "send" next to it?

MEMBER WEINTRAUB: No.

DR. ROILAND: Bill, if you are comfortable with saying it over the phone you can do that or you can email me.

MEMBER WEINTRAUB: Well, let's see if this works.

DR. ROILAND: Okay.

MR. SINGH: Does everybody who has a clicker showing like a number on the remote when you hit it?

MEMBER NEEDLEMAN: Mine didn't.

CO-CHAIR ASPLIN: I just want to

reassure everybody, we'll pick the pace up as we go on.

(Laughter.)

CO-CHAIR ASPLIN: This is only going to happen once.

MEMBER WEINTRAUB: Rachel, did you check? Did you see the chat response?

DR. ROILAND: Hi, Bill. Yes, this is Rachel. I have your vote on the chat response. And I also have a vote from Andrea Gelzer and Troy. And Mary Ann, I have your text. Yes, I have your vote as well.

MEMBER RAMBUR: And do you have Betty's? I sent mine via text and email.

DR. ROILAND: Yes, I got it.

MEMBER RAMBUR: Thank you.

MR. SINGH: And, Dr. Ratliff, I also captured your vote, too.

DR. ROILAND: And, Larry, this is Rachel. I got your vote as well. So we should be all, so we should be all good to go. Sorry about that.

MEMBER BECKER: Terrific.

MR. SINGH: So it appears that the voting platform is good to go. We've got everybody situated and all the clickers are going. So now we are going to be commencing our official vote for Measure Number 1598, Total Resource Use Population-based PMPM Index on the first criterion on a high priority.

Your options are 1, high; 2, moderate; 3, low; and 4, insufficient. Please vote now.

(Voting.)

MR. SINGH: Everybody in the room and everybody on the phone -- or everybody in the room, can you please submit your vote one more time.

(Voting continues.)

MR. SINGH: One last time.

MEMBER WEINTRAUB: Did you get everybody on the phone? Do you want us on the phone to do it again?

DR. ROILAND: No, we can resubmit your votes. Thank you though.

MEMBER WEINTRAUB: Okay.

MR. SINGH: Dr. Weintraub, can you please submit your vote via your proxy? Thank you.

MEMBER WEINTRAUB: Oh, you want me to do that rather than the chat now? I'm still, I don't see email from a proxy to respond to.

MR. SINGH: The chat can work, too.

MEMBER WEINTRAUB: All right. I will do it again on the chat. All right? We'll try it again. Please let me know if you get it.

DR. ROILAND: We have your vote, Bill. Thank you.

MEMBER WEINTRAUB: Thanks.

CO-CHAIR ASPLIN: We're just sorting through the proxy issues here with the number of people we have on the phone to try to get the right denominator. Once we get rolling this is, this is going to be much more efficient. This is not going to be a Chicago-style vote, however.

DR. ROILAND: Can we do it one more time? And if not, we can maybe move to a hand

vote.

CO-CHAIR ASPLIN: I think it's worth the time to get this because it's going to be really cumbersome to do anything else.

MR. SINGH: So, we're at 19 now. Okay, so voting is now closed.

The results for --

CO-CHAIR ASPLIN: For clarity, we're voting on high priority, last time. Those of you on the phone, we already have your votes.

MR. SINGH: Okay. So, all votes are in. The voting is now closed for priority one for high priority for 1598. The results are 100 percent high.

MEMBER MARCINIAK: That wasn't my vote. I'm from Chicago. I feel very comfortable with that.

(Laughter.)

MS. O'ROURKE: For those of you on the phone, we're just discussing in the room how to proceed with voting. So apologies for the delay.

Why don't we just, if everyone is

comfortable, take a vote by hand on this criterion just to get us back on track. And then we'll try the clickers again for the performance gap vote.

So, people on the phone, we have your votes. If you don't want to change, no action needed for you.

In the room, if you could just raise your hand as we go through.

Well, first let me say, does anyone mind voting publicly? Okay, so if you, for high priority, if you wish to vote high, please raise your hand now.

(A show of hands.)

MS. O'ROURKE: Okay, we have eight in the room.

If you wish to vote moderate.

(A show of hands.)

MS. O'ROURKE: We have two in the room.

And if you wish to vote low.

(A show of hands.)

MS. O'ROURKE: We have one in the room.

When we add those to the phone votes we have our correct total?

DR. ROILAND: So we have 16 for high; 2 for moderate; and 1 for low. And that equals 19; right? Okay. So that's the vote for high priority for Measure 1598.

MS. O'ROURKE: And we apologize for that. We are going to get you back on conversation while we fix the voting software in the background. So apologies. Thank you for your patience, everyone.

CO-CHAIR ASPLIN: Very good. Thank you.

So, we are going to move on to the performance gap. And we will also be voting on this gap here, an opportunity for improvement. And before moving to the vote, are there any additional comments? We did discuss this a little bit. Are there any additional comments on the performance gap in this measure, an

opportunity for improvement?

(No response.)

CO-CHAIR ASPLIN: None in the room. And I'm not seeing any online.

So, if we could move ahead to -- Are we ready or do you want to do another manual vote?

MS. O'ROURKE: Let's try the software. And if it doesn't work the first time we'll just move to a manual vote and get someone up to reboot it for us.

So, Irvin, if you could read the platform.

MEMBER WEINTRAUB: So, we should go ahead and use the chat function?

MS. O'ROURKE: Yes. If you are online, please go ahead and chat your votes to us.

In the room, if you could use the clickers.

MR. SINGH: Okay. So we will be voting on the second criterion and that is gap in

care/opportunity for improvement for Measure 1598. Your options are 1, high; 2, moderate; 3, low; and 4, insufficient.

Please begin voting now.

(Voting.)

MR. SINGH: This one looks good.

So, the results for the second criterion gap in care/opportunity for improvement is as follows:

We have 8 votes for high; 10 votes for moderate; 3 votes for low; and 0 votes for insufficient, which equals out to 42 percent high; for number 2, 53 percent; for number 3, low, 5 percent; Number 4, 0 percent for insufficient.

CO-CHAIR ASPLIN: Very good, thank you.

And we will move on to the final vote in importance which is measure intent. Are there any comments prior to the voting? If not, we can read the slide. We'll have some additional information about the specific nature of the

vote.

Let's move on to measure intent.

MR. SINGH: Okay. So we are voting on the third criterion, measure intent for Measure Number 1598. Your options are number 1, high; number 2, moderate; number 3, low; and number 4, insufficient.

Please vote now.

CO-CHAIR ASPLIN: For those of you on the phone, the description of this third vote is the measure intent of the resource use measure and the resource construct is clearly described.

(Voting.)

MR. SINGH: Okay. So the results for measure intent is 12 high; 2, moderate, and 6 votes; number 3, low, 1 vote; and there were 4 insufficient.

So just to reiterate, for 1 high, there's 12 votes; number 2, moderate, there's 6 votes; number 3, low, is one vote; and number 4, insufficient, 0 votes.

CO-CHAIR ASPLIN: Very good. So those

are our votes on importance.

And we can move on to scientific acceptability. And our lead discussant for the first section of scientific acceptability is Herbert Wong. Herbert.

MEMBER WONG: Very good, Brent.

So, I'm thinking about for the purposes of efficiency, talking about the scientific acceptability both in terms of reliability and validity all in one conversation because they often kind of spill over into these different characters. When we do the voting itself, I think that we can go systematically through this.

So, you know, just as a reminder for folks that the numerator for this particular measure is a risk-adjusted total resources that are associated with treating our members. This includes professional facility in-patient and out-patient. It includes pharmacy, labs, radiology, ancillary, and behavioral health.

The denominator is really a care group

average. Right? And it's designed to be somewhat flexible, depending on what the -- whoever is deploying it is. So you have a numerator aligned to values and you have a peer group.

There are a number of exclusions that are included in this particular metric. They're excluding services that are not office-based. They are excluding convenience care clinical visits. They exclude providers that are not a physician, physician assistants, and nurse practitioners.

They look at service lines for specialty-based services, practice, and specialty.

They include the following specialties: family medicine, internal medicines, pediatrics, OB/GYN, and pediatrics.

There are a number of other adjustments. Basically member aged over 64 are excluded; members aged -- under age 1 are also excluded.

There is a criteria there that the enrollment has to be more than 9 months, basically excluding anyone who has less than 9 months, up to a year's time.

Members who are not attributed to a primary care provider is also excluded.

One of the things that was mentioned earlier on was the dollars per member has to be -- if it's above $125,000, those components are excluded.

The risk adjustment is designed for age, gender, and diagnoses.

In terms of the reliability testing, when they first started -- when this was first approved they basically deployed three different methods in terms of looking at reliability. They deployed 90 percent random sample, a bootstrapping technique, and the analysis of performance over time.

The data that they employed was from health partners looking at the years 2007, 2008, 2009, for 19 individual providers in this

particular area. And, roughly, the members in -- for each of these years was roughly 250,000 to 300,000 members.

In terms of the new data, looking at reliability they looked at I believe it was the years 2014 and 2015, and they had roughly 470,000 members involved there.

So, in terms of reporting the reliability, they honed in on the bootstrapping and 90 percent sample. Over all, the results are consistent across -- are consistent within a group. There is some variation across the group.

So, in general, the assessment from both the staff here at NQF, and members' preliminary analysis of that, kind of gave the assessment in terms of reliability over all, in terms of the testing, as high. And that's in generally I'm speaking.

In terms of scientific validity, again the data source that they deploy were the same years that they used for the reliability, again focusing on health partners' commercial

population again, with roughly 470,000 members.

They examined correlations between measure components and measure scores with other markers of utilization.

There are a large number of tables that I'm sure that folks have kind of reviewed there that look at these correlations. I won't go through them all there.

Here are some of the take-aways I think I have looking at both the reliability and the validity. I think that in general I think that committee members, at least in the preliminary stages, looked at reliability and kind of rated it as high over all.

When they looked at validity, I think that folks kind of rated that in a more moderate level a little bit.

The big concerns that kind of emerged from this conversation were, one, concerns about attribution. The second concern was a concern about some of the conceptual bases of trying to pull this metric together. Namely, why is it

that one used a 9-month cutoff as opposed to something else?

There is a notion of generalizability that much of the testing was done in a localized area. And so there was some question of whether or not this particular metric is applicable on a nationwide scale. And then there were also some comments about how looking at some of the specialty groups that were relatively narrow, looking at some of them, but didn't include other specialty groups.

My own general comments about this is this notion that when you look at this particular metric I think that one has to be a bit cautious, in particular looking at the peer group. So you do have a denominator. You do have a -- you have a numerator of a group that you want to compare. You have a denominator where you want to benchmark it to another group.

There is some notion there from a conceptual point of view of how you could have some flexibility of choosing that. So, for

instance, if it's cutting across a geographic space, if you have a large health plan and you're looking at this, and how you choose your denominator, your peer group becomes important.

So, for instance, if you have two different locations, one is, say, New York and one is in Florida, if you don't choose the right peer group what's -- who's to say that you have the right metric, that is the right average, unless you thought all about how one is basically doing that?

There the exclusion of $125,000, there were some comments, I think, that reviewers had of whether or not that was a bit arbitrary. And I know that it was well-established back in 2012 that the developers basically increased that rate from 100,000 to 125,000 to reflect some inflation sort of basis there.

Let's see. I think that for the 9-month period some additional sensitivity analysis could have been done to kind of demonstrate that this is not, should not be a concern, whether or

not you remove some of that restriction, some comparison groups there.

So, I would say that over all the approach, both in terms of reliability and validity, are pretty well-accepted methods in the literature and that the approach is pretty solid in terms of how they approached it.

I think that some sensitivity analysis might be in order for some of these parameters.

I think that some of the conclusions that one can come up with is that for these particular metrics is is generalizable within groups, but one has to be cautious when looking at it across groups.

So high level kind of comments on both reliability and validity. I'm sure that the committee members will kind of add their two bits to it as well.

CO-CHAIR ASPLIN: Thank you, Herbert.

So, we are going to address and assess reliability first. And then we will have the conversation around validity and several things

that I have heard mentioned here around attribution, truncation changes, exclusions, et cetera, will be talked about in the validity section.

MEMBER WONG: Yes.

CO-CHAIR ASPLIN: Any specific comments on the reliability? Sue, do you have anything to respond initially to on reliability before we get the committee comments? Or would you just want to wait for committee members?

MS. KNUDSON: You know, I think the most general comment I'd have -- and I'd ask Gary and Chad, too, because we are all intently listening -- is so it's our understanding attribution, with this round of review of our measures, continues to be guidelines. So we have offered the attribution method that we have used within our health plans.

What I'll add onto that is in our work across the state with Minnesota Community Measurement we analyzed different attribution measures. So, first-off, the inclusion of the

specialties in ours, what works, that works in our market. And that is widely accepted by our group.

And so when we do apply the measures, either statewide like we did, or nationwide, as that comment was made, it is important to have an agreeable attribution that we were able to derive to that. What we found here in this state was, you know, most of the models are returning the same types of results. So, it's as much about the inclusion of a type of providers as the culture that you're operating in.

So that would just sort of be my general comments on attribution, which I think is -- follows both of these sections. And then the other comments about the peer group, the truncation limit, and the 9-month continuous enrollment we'll handle in the reliability.

CO-CHAIR ASPLIN: Okay. Comments on reliability? Cheryl.

CO-CHAIR DAMBERG: Thank you. I appreciate the information provided on the

reliability in terms of consistency of results, test/re-test. But I don't think I saw in the materials any signal-to-noise test to be able to look at whether you are able to discriminate the performance between providers.

And that was concerning from my perspective, given that this type of measure is frequently used in the context of value-based payments.

MS. KNUDSON: Do you want us to handle all the reliability questions now, Brent?

CO-CHAIR ASPLIN: Yes, that would be great, Sue. If you could respond to that, that would be great.

MS. KNUDSON: Okay. We'll start with that signal-to-noise one. I'm going to have Gary respond to that.

MR. KITCHING: Sure. In that paper, when we ran our paper we looked at the bootstrap analysis. And what we did was we compared the bootstrap analysis, or the result, to the actual result. And that's kind of the noise, if you

think of it that way; so the differences within a provider group.

When you compare that to the what you're trying to measure against, so that's across all provider groups, and this is the spread between each provider group is about 110 percent. So the signal-to-noise issue, what we're looking at is less than 1 percent would be the noise. And the signal then is all the 110 percent.

So, really, if you look at a provider group that has a range of, say, 0.9, the noise in there is going to be, like, 0.89 to 9.1, or 9.1. But the spread that we're talking about is over the course of, you know, 100 points. So you'd be able to confidently say that that person, the ranking we gave them is their result. And I think that's what we're, what we're trying to conclude, or that's what the paper concludes. And that's what that testing mechanism does.

We've also done the reliability, which I think you're referring to, which is I think

people actually try to think about that 0.7 reliability stat. We've done that test as well, and it's on our website. And we're actually just performing that now. And we'll have it up there soon.

Those results have the same, conclude the same results, but that 0.7, if you have a 600 and above population size, that returns back a reliability score that's positive as well.

CO-CHAIR ASPLIN: Thank you. Jim.

MEMBER NAESSENS: Yes. Actually, most of my comments are more about the attribution methods. And as a result of attribution in this whole process we've experienced in Minnesota Community Measurement, you know, experience where there's actually a little bit different attribution applying the same total cost of care and research. And in that attribution process it turns out that they used similar definitions of specialties, really focusing, understanding as an intended frame to the primary care practice.

But in the way that the attribution is

applied we end up identifying lots of patients who see many nurse practitioners and physician's assistants in specialty clinics. But because they're labeled all physician assistants and nurse practitioners as primary care providers, we see many referral patients being counted and captured in our system and attributed to us as primary care patients and being included in these measures.

So, for example, only 3 percent of our primary care practice gets picked up because it's only a selected care group. But, also, 60 percent of the non-primary care patients who get captured have cancer in the care patients, which will really skew the information.

So, if we're accepting a method and saying that it's valid without an actual -- with only a recommended patient attribution method, how do we say that -- or what, what sort of standard does National Quality Forum have to put on a measure whose underpinnings are not really valid in terms of patient inclusion?

The other comments, while I just have the floor, would be some of the things that Herb mentioned as exclusions are actually exclusions in their attribution, not exclusions in their measures. So the only exclusions, it's primarily who's between that age of 1 and 64, outlier patients get truncated, they don't get excluded. And the threshold, if the intent of eliminating outlier patients is to make a more normalized distribution, it still leaves a chunk of patients about 20 percent higher than the mean in the distribution.

In our experience with the data across Minnesota we have 5 percent of the whole population being outliers, which doesn't normalize that distribution very much.

What I would recommend that we do in terms of approving the handling of outliers is to actually exclude them. Just like the Medicare spending per beneficiary excludes patients who are transferred between systems, in this case many of those outliers are actually not being

handled by the primary care group who is monitoring and managing most of the patients. The specialty practice is only seeing part of the care, and those patients would also be best to be just left out of the analysis rather than skewing information. They'd be handled in a very different fashion.

CO-CHAIR ASPLIN: All right. So a couple different issues. Taroon is going to the first part of your comment, Jim, relative to the NQF's position.

DR. AMIN: So, and there's many of you on the committee that participated in the original construction of what we discussed in terms of the measure specifications. And this is what actually is going to key up for conversation later this afternoon as it relates to our attribution panel.

To address the attribution question in particular, as we looked at the measures, as we looked at cost and resource use specifications, originally this standing committee, former

members but the standing committee, outlined there are several components that could be submitted as specifications for guidelines. So, by nature, some elements of measure submission while they're important to understand for context, were not part of the measure specifications that would be under review for this committee.

And one of those elements is the attribution model. That certainly, so that is the current construct of what this committee should evaluate this measure by.

I think, Jim, you're pointing out a really interesting challenge with that construct which, obviously, has been challenged by our Attribution Committee, which will be part of our conversation later today around whether that needs to be not just a guideline but essentially a specification in the measure. However, that is not what is in front of the committee at this point.

The committee's current guidance to

the developers and current evaluation process still allows that to be a guideline. I think you will, each of you will have to struggle with how to interpret that in the context of the validity and the reliability of the measure specifications.

And I will keep the second point around exclusions up for the committee discussion.

I don't know if that answers your question or may or may not be satisfying, but that is currently what the evaluation process that we've laid out is structured as.

CO-CHAIR ASPLIN: Thank you, Taroon.

So, my sense is that we're going to have more issues to talk about on the validity side than the reliability here. So I'm just going to ask Jack or Srinivas, did you have reliability comments or questions? Jack, go ahead.

MEMBER NEEDLEMAN: Just very quickly. I want to reinforce what Cheryl said. And I

think Jim's comment underscored the issue.

The data that's available for constructing these measures is at the individual patient level. But the unit of analysis is the group or the medical group or what the -- or the individual provider, whatever the aggregate is that's actually in those reports. And the reliability measures should focus on what's happening at that level. And, frankly, I'm not seeing very -- I'm not seeing as much as I would like to see to really evaluate the reliability there.

A lot of reporting of what the population, stability at the population level, but not a lot of reporting of the stability at the group levels and the intergroup levels. And it's very had to assess how reliable this is in practice as opposed to how reliable it is at the population level, where it looks very stable.

CO-CHAIR ASPLIN: Sue.

MS. KNUDSON: Yes. I just want to clarity, Brent. Those are our group. That is

the group-level testing.

MR. KITCHING: Population level.

MS. KNUDSON: Yes. Population at the group level, so 66 groups.

MEMBER NEEDLEMAN: Yes, but just to be clear. What has to happen is we've got to be looking at what the averages are across the groups, not what the averages across the population and how stable. A 90 percent sampling from, or bootstrapping at the population level is not the same thing as getting a separate set of bootstrapping or estimates of the stability of the measure group by group and looking across the different groups at that.

MR. KITCHING: Yes, this is Gary again. Yes, we did the -- the sampling is actually within the group. So the random sample that we take, with replacement for the bootstrap technique, is within the 66 groups. So in essence you sample them 500 times within that whole, within the group. And then you compare that end result to the actual. And that's how we

know that the resulting variation is within that particular group.

There again it's across the entire 66 population of groups, that's the 110 percent. So we're confident that we know the signal-to-noise ratio and we can tell the difference between performance for each of those 66 groups. So it is a group analysis.

CO-CHAIR ASPLIN: Are you satisfied, Jack? Or this seems like an important -- because I, I interpret it, just like they said, that the testing was at the group level. So maybe I want to make sure we don't have -- we're not talking past each other here, that we all agree on what was done. So, does that make sense, Jack?

MEMBER NEEDLEMAN: It makes sense. Again, the relevant number here is not those .00 -- I forget, there was a third zero before you get to the first figure number. But that 110 percent variance. And that's got to be compared against the variance across the groups to figure out how stable and reliable the measure is. If

we're trying to measure intergroup comparison, 110 percent actually looks large as the degree of uncertainty around what's a relatively small group variance in the RCI measure.

MR. KITCHING: Yes, I think, just in layman's terms I'll just try and communicate what we're trying to test here.

So, if we take any one of those provider groups, the resulting score, those three bars, if you look at the graphic, they're pretty tight. There's no difference between the index scores for all the different testing techniques that we put in place. So, for any different group there's no difference in the results, depending on what sampling we did for it.

The 110 percent it's at the range of performance across our whole network. So we have some provider groups that are performing at 70 percent of average; other provider groups that are at the 1.8 percent of average, so 80 percent higher than the average.

So the difference between one provider

group that's performing really well -- 70 percent of average; another group of 80 -- 100, yes, 80 percent above average of 1.8; that's actually the range we're measuring. So we can tell that provider group that's at the very best performer or very worst performer, or highest performer, is that high at 1.8.

MEMBER NEEDLEMAN: Yes, but if you pull into the 10 to 90 percent range as opposed to the extremes you get a much, much smaller variance than that. Much smaller.

MR. KITCHING: Well, I mean obviously it's 10 percent of average and 90 percent, so, yes, I would agree. There's still a 1 percent variation in that performance score. And that's still vague. I mean, there's still 20 points between 10 percent above and 10 percent below average. And you're still looking at that 1 percent variation in that person's score. And he's not going to go from 10 percent below to 10 percent above, he's only going from 10 percent below to 9 percent below. And that's the noise.

MS. KNUDSON: And as I mentioned before, 1 percent is still meaningful.

CO-CHAIR ASPLIN: Srinivas.

MEMBER SRIDHARA: Thank you. So, I wanted to just comment a little bit about sort of practical implementation, so just for context as I mentioned before. Prior to my current role here, Advisory Board, I was the Maryland HCC director, and we tried to actually implement this measure set along with several other organizations as part of an NHRI Robert Wood Johnson Foundation.

So, just in terms of some of the things that we're talking about, whether the attribution being guideline instead of a specification, two comments on that: one from having sat on the Attribution Committee; and two from sort of the actual approach here.

These sorts of flexibilities, be it the attribution approach or the local reference populations, actually make adoption more likely. And including sort of some conversations around

how you might define primary care. All of these and how someone might go about implementing it actually are important in getting it through and actually using it, convincing the provider groups, tuning it to your local population and making that work.

That said, that makes it harder to sort of say how do we measure reliability across nation due to, you know, sort of external validity that was just discussed as well. So, you know, you wish there was a national benchmark to compare it to and make that easier to do, but in terms of comparing local provider populations, usually, at least we found that you had a better success, say, comparing to other providers that are under the same payment models or related factors that are going on.

So in that way in sort of statistically performing it and doing it and convincing folks in a practical side, it made sense. So, I think there is a tradeoff between getting very explicit and specific on how you do

every element of it, which helps us do the scientific study of it. But in terms of application, I think even if you endorsed it that way I think individual folks or groups who would implement it would have to necessarily tune it to make it applicable in their population.

I think on the attribution part, you know, we've discussed some of that. And I think we have a whole lot of debate in the Attribution Committee around whether it should be a fundamental unit of, you know, the specification or is it flexible? And I think, again, in terms of evaluation it makes it easier if you have a -- you know, you need the attribution as context, at least for the evaluation, which we've talked about here. But I think, again, in terms of applicability it's you would want the flexibility.

The second is some of the concerns around, say, the 9-month, you know, period of enrollment, et cetera, these are to me things that are relevant for commercial population. And

then I think it makes sense if you thought about what is the average enrollment period and sort of if you look -- you know, and people do, actuaries do PM calculations and comparisons across commercial populations, 9 months is probably a pretty good number, if you think about it. It's different, say, if you did Medicare, and certainly very different if you did Medicaid.

And so I think that's, that's an important thing to keep in mind here. You know, something that I wish this group had taken on, I know that was mentioned already about lack of that kind of data in their population, but thinking about how to apply it to Medicare or Medicaid, they've done some work to think about it. But I think figuring out how this might apply in those populations would be important and specific guidelines.

I just wanted to provide that for context.

CO-CHAIR ASPLIN: All right, very good.

We are going to move on to the vote on reliability. Irvin.

MR. SINGH: Okay. So we're going to be voting on the reliability criterion for Measure 1598. Your options are Number 1, high; Number 2, moderate; Number 3, low; Number 4, insufficient. Voting has begun now.

(Voting.)

MR. SINGH: Okay, so all votes are in. Voting is now closed.

So the results for reliability for Measure Number 1598 are there's 10 votes for high; 7 votes for moderate; 1 vote for low; and 1 vote for insufficient, which accounts for 53 percent high; 37 percent moderate; 5 percent low; and 5 percent insufficient.

CO-CHAIR ASPLIN: Thank you.

We then turn it over to Mary Ann Clark who is the lead discussant on the validity section of scientific acceptability. Mary Ann.

MEMBER CLARK: Yes, hi. I guess I was under the impression I was doing the 1604

reliability and validity, so I kind of focused on that.

CO-CHAIR ASPLIN: Well, Herb sort of covered both here. Maybe --

MEMBER CLARK: Yes.

CO-CHAIR ASPLIN: -- I have it wrong in the documentation. So that's, that's fine. We can just open it up. We've already raised many of the validity areas that committee members want to address.

MEMBER CLARK: Right.

CO-CHAIR ASPLIN: So --

MEMBER CLARK: Right.

CO-CHAIR ASPLIN: -- that's fine. We can table your comments till 1605 then. That's okay.

Yes, Janis?

MEMBER ORLOWSKI: Based on the discussion where a number of the reliability issues were raised, the vote is surprising. And I think it would be helpful if, it would be helpful to hear both sides of the story. You

know, again, the vote does not reflect the discussion we had. And that makes it difficult to be able to hear both sides, is my comment.

CO-CHAIR ASPLIN: I guess it depends on how you interpret the conversation. Because I -- Is there a concern though that you think was not being addressed in terms of questions about reliability?

DR. AMIN: Brent, maybe I can help. Maybe I can help.

So, the conversation around reliability focused a lot in terms of -- and Jack, I'm going to, Jack and Herb, I'm going to ask you guys to weigh in here -- it sounds like there was a lot of concern around the reliability testing approach that was used by the developer. And I think the question that's being raised is did, generally -- it sounds like the committee, while it had some concerns around the reliability testing generally still moved it with a high rating.

And I think there's a question of, you

know, can we have a little bit of discussion around everybody sort of reconciled that.

Is that a fair question that's being raised by a committee member? So given, Herb, your discussion, and Jack, that you had raised this concern, maybe having a little conversation around how you weighed the testing here, not necessarily -- and others who may agree to that.

MEMBER WONG: Right. So I'll, I think I'll start it off or at least kick it off.

I think that, you know, from -- I think that part of the thought process here is to think a little bit about the measure from two aspects of it. One is a conceptual basis, and the other is from the testing aspect of it in terms of the reliability. I think that part of the conversation had drifted towards a conceptual basis of that. And some of these components, whether, you know, one of my comments, for instance, is looking at the denominator and making sure that we have the right baseline and things of that nature. That gets into a space

that's a little bit, I would say, conceptual. Right?

And there is some flexibility that is basically built into that.

When you go into reliability in terms of looking at the data of what they have and what they have done, the methods are traditional, standard. The approaches are really out there, as has been well known, it's well documented how you do those sort of things. And they have done that.

And the, in terms of the findings on it, you know, one of the things I characterized from taking a look at the tables on that in terms of what was reported out, that in my judgment if you take a look at those correlations and things of that nature, it seemed to be performing better when you look at it from a group perspective. But if you look at it from across groups there was greater variation. There was greater concerns there.

So, you know, in that judgment, you

know, from my thinking of it, it would be in either the high or moderate space over all. So trying to separate some of the concepts versus what they had deployed. So, you know, high level kind of characterization there.

I'm not sure if Jack has anything to add.

MEMBER NEEDLEMAN: Yes. Very quickly.

I, the comments of the developers got me back looking at the documentation, and on 43 of the -- page 43 of the PDF there is actually a group by group description of the results of the reliability testing. And I went back to look at that.

It's not clear from the earlier summary description, but I did go back and look, and the variance at the group level in these measures between the 90 percent sample, the bootstrap sample, is in the third decimal place of a measure that's measured somewhere in the, you know, where the relevant number is in the first and second.

So, I, having looked at that data I am now comfortable the reliability of the measure is fine. We can discuss the validity a little bit more.

CO-CHAIR ASPLIN: Thank you both for jumping on that. Thank you.

All right, so I have heard a few different topics come up. Attribution is really a -- we could talk about that from a validity standpoint or usability. And it sounds like that's going to be one, another one of those issues that will be difficult for us to fully resolve, given the current guidance to developers and the charge of this committee today. And maybe another moving target in terms of the overall endorsement process over time.

Exclusion, truncation, and then, Janis, I believe you had some comments earlier around. that came up in the importance dialog around sociodemographic factors. Do you want to make those comments now? Great.

MEMBER ORLOWSKI: I would say that my

comments fall into two specific areas regarding the risk adjustment for sociodemographic status.

First is that I believe that the two tests that were done are inadequate. I believe that taking a look at income and poverty level at a 5-digit ZIP Code is an inadequate measure. And there's been much reported about this particular thing.

I also find a black versus non-black again surprising and unreliable. And so I do not believe that there has been adequate assessment of disparity.

The second comment that I would make is that the developer says that by using these two measures, both of which I already said I believe are inadequate, I see no more than X percent movement. So that in one case 97 percent of the hospitals had no movement. The other 95 percent. And when you look at it at the level of the hospital, I would argue that you are looking at it to say that there is minimal movement, that that may be an appropriate statement.

And so, for example, 5 percent of the hospitals across the United States are large academic medical centers and they care for 25 percent of Medicare patients and 40 percent of Medicaid. So, so to have a 5 percent of hospital movement you'd have to say which hospitals.

Are urban hospitals affected by this movement? Are rural hospitals? And I don't believe that I have enough information to be able to take a look at those particular variables.

And so I would say, and then I think the third thing is is that the -- we've already talked about this before -- is that for the majority of this measure you're bunched between 10 percent and 90 percent. And I'm not sure that these measures help you look at variations in SDS within the population.

CO-CHAIR ASPLIN: For clarity, some of, the middle section of the comments sounded like it may have been related to the 2158 measure, the Medicare Spending Per Beneficiary, since this is not a hospital accountability

measure?

MEMBER ORLOWSKI: Oh, you know what, I'm sorry. I --

CO-CHAIR ASPLIN: Okay.

MEMBER ORLOWSKI: -- was reading from my -- that's right.

CO-CHAIR ASPLIN: Okay. Now, other comments on validity? Or do we -- Jack, go ahead.

MEMBER NEEDLEMAN: Yes. I have a question about measure construction because we can't talk about validity until we know what's in it.

So, standardized pricing model, and I'm curious to know how, what standardized prices are being used for the hospitals. Some hospitals, some insurers that pay on the basis of charges or discounted charges, still a fair number still do it that way. Medicare pays on the basis of DRG. So I'm wondering at what level is the hospital visit priced? Are we capturing the differences in variations in charges for

hospitals? Are we simply looking -- and this is both for the standardized measure but also the -- it will be more important when we look at the second measure in this group, the pricing, the prices paid measure.

So what's the level at which the hospital prices are standardized and aggregated? DRG? Individual charges? Individual diagnosis? I need to -- it's not, I did not see it specified here.

CO-CHAIR ASPLIN: Sue.

MS. KNUDSON: So, two things. First-off, I just want to be crystal clear on the previous commenter before Jack, Brent. We did SDS testing but that is not the SDS testing as described, that was our testing. So, I think those comments do apply to the other measure and not ours. So, we'll get to that because that is not descriptive of the testing we did.

And then on this particular question about the standardized pricing, Chad will be taking that question.

MR. HEIM: Yes. So just to give a real brief overview of how the key CRVs are calculated, again, our approach hasn't changed since the last endorsement. Key CRVs, which is the resource use that we are using, base is a linear scale where we are doing a patented approach for assigned resources to draw healthcare services. So we are getting that relativity within using well-established CMS weight at the starting point to get the relative event.

And specific to your question, Jack, we are using MS-DRGs for that relativity using CMS weight system there.

And then to get the relativities across in-patient to out-patient to professional legal through a repricing scenario to get that relativity going across.

CO-CHAIR ASPLIN: You have a follow-up, Jack?

MEMBER NEEDLEMAN: Yes. Not on that but a second question related to the

construction.

You mentioned in the documentation where drugs are carved out you're putting in the average drug price across those. You said and other carve-outs we deal with the lowest common denominator. One of the most common carve-outs is mental health/behavior health services. And it's not discussed explicitly.

So, what percentage of your groups have mental health or behavioral health carved out? And how precisely are you dealing with those carve-outs in the construction of these measures?

MS. KNUDSON: That's a good question, Jack. Glad you raised that.

You know, Minnesota has parity with mental health benefits for years. As an organization and one of the top providers of mental health services and leaders in the state, we are quite committed to both covering as well as caring for patients and members with mental health conditions. And so we have 100 percent in

our data.

We were just simply making that notation because we know that in other markets it could be, in historic data, carved out at higher rates. But we do not have that issue, it's all in.

CO-CHAIR ASPLIN: Any other -- Well, one quick question I had, Sue. Are you aware of users that are -- this is obviously developed using the Hopkins ACG risk adjustment tool, obviously a widely used commercial risk adjustment methodology. Are you aware of anyone using a similar approach with different risk adjustment tools?

MS. KNUDSON: Yes, honestly that's something that we talked about in the first endorsement process, for some of the committees who remember that dialog. Yes, I think, so for example, different states might have a lot of payers or providers who are used to the Optum ERG risk adjustor. And it's something their communities invested in, each organization

individually.

So I know lots of folks are maybe using our specs but applying a different commercial risk adjustor. I think the bottom line in our testing around that was that for the commercial population, given the disease prevalence, et cetera. And ACG is the one we used, but recognizing other markets use other commercial risk adjustors and they're applying them.

We do let them know that it's endorsed with ACGs but I think it goes to their uptake and what they're able to deploy quickly in their marketplaces.

CO-CHAIR ASPLIN: Thank you.

Toy online and then Nancy in the room.

MEMBER FIESINGER: This is Troy Fiesinger.

I appreciate your explanation about SES and risk adjustment. I am still concerned that if we include and, if possible, come up with a test model in other diverse communities to see

if your data still holds true. It's clear that there's going to be very large disparities in income and cost of care/quality of care. And we have patient net systems working hand-in-hand with commercial systems.

But it would help to have more reassurance that's been tested in multiple communities.

MS. KNUDSON: I'm not sure I understand that question. But just here's what I know.

In our SES testing we tested socioeconomic status variables and discussed those. We did not test race, ethnicity, and language per the NQF kind of position on that. We take an approach as an organization, we are very committed to health equity. And so if we talk about race, ethnicity, and language, or REL, for our quality measures, for example, we use a segmentation approach and identify gaps in care and work to close them.

But we did not do that testing because

it was specific around SES, which is different variables, income, education, et cetera.

So, Taroon or one of the NQF staff, I'm not sure if maybe you want to comment on that direction?

DR. AMIN: I think we had a little bit of conversation around this SDS question. I think what might be helpful is to just if you help our understanding, if you don't mind just reviewing what you did in terms of SDS, just so that we're all on the same page in terms of methods that you used and the variables that you tested, just so that we're all working off the same songbook. And then everybody in the room, I would just encourage you to take a look at the PA as we are talking about this and just make sure that we're all on the same page because it's, obviously, an important question for everybody and we just want to make sure we all understand what was done here.

MS. KNUDSON: Okay, yes.

DR. AMIN: Is that a fair ask?

MS. KNUDSON: So Gary is going to go over our approach to socioeconomic testing.

DR. AMIN: Thank you.

MR. KITCHING: Okay, yes.

So what we ended up doing was we took two different data sources for our SES. We have this one that's the census, that's at the track level. Again, just for informational purposes, a track level is at, you know, 1,000, an average of 1,000 patients up to 8,000 patient blocks. Not any ZIP Code level or by ZIP, but actually a little bit more precise. But still, you know, a good swath of folks.

And then we have another database that we purchased that's actually a commercial database that has household-level data. So we did the testing at both those levels just to kind of verify. We know that people are going to have that additional, more precise data set, so we wanted just to make sure we test the census data and then we go back and actually test on a more precise data set to see if we see the same

findings.

And so we did that testing. We just ran a, you know, a simple regression analysis. I say simple, but obviously there's a lot that goes into that. But we took a look at the income levels. So we took the average of the income levels that we have and use those tracks, put that into the model. We took the ACG score, we put that into the model -- clinical risk adjustment.

And then we both put the commercial population and then we expanded it out to the Medicaid population because we wanted to make sure we cut across to show that there is some variation when you do include low income folks. And we're saying that the Medicaid folks are going to be at lower income than the commercial folks.

And so when you put all that in there, into the model, and then the table that we have displayed shows you what factors are driving the end result. And so then PMPM, the risk

adjustment PMPM, and the resource use would be the TCRV PMPM.

And so you can see in the one table already produced the top line, $0.13 is going to be the total reimbursement attributable to the 1 percent increase in income. So that's kind of how that would bear out.

And then you look at the ACG score, it's $4.22. So there's obviously considerably more association with the clinical risk than there is with the income. And then, furthermore, you go out to the commercial risk and Medicaid and you have $133 in that one example there.

So you compare $0.13 for income versus $4.22 for ACG or clinical risk, and then the payer market or the payer delineation of Medicaid to commercial is $133.

So, obviously, those factors are a lot more substantially influential on the end result, ACG and the payer, than there is on income. So we were able to include that income really has no discernible variation or driver on a provider

score beyond when we already control for the ACG and the clinical risk and the payer.

So when you look at it that way, we concluded, as I think most of you will, that when you adjust for the ACG or the clinical risk and the payer, you're controlling for most of the variation that's going to come from any kind of an SES analysis or any SES adjustment that would be needed.

So that's kind of the approach we took. When we tested it with the Experian data, which is again the household level data, a lot more accurate information, the same results bore out. And, actually, the income had less of an impact, to be honest.

MS. KNUDSON: And I would just note that Gary has said in that opening comment that we tested both sources we had available to us, but understanding from an uptake point of view, census track is most available. But because we had that secondary commercial source for our other business purposes, we wanted to do a more

granular test to more inform the outcome of this work.

MR. KITCHING: Thank you, Sue. Yes.

CO-CHAIR ASPLIN: Thank you, Gary. Thank you, Sue.

Nancy.

MEMBER GARRETT: So I just wanted to comment, a couple of comments on the SES adjustment that this is.

So, first of all, I think it's really great that they went to the census tract level. A lot of the work we're seeing in some of the Medicare measures has been at the 5-digit ZIP Code level. And as Janis mentioned, that is really, a lot of research is showing that just doesn't get precise enough to start to attribute to individuals. So I think that's really helpful.

And the extra step of looking at the Experian data also I think is really interesting. One caution I would give there is that we've been trying to understand better how that, those kind

of commercially available data sets can be used as a SDS indicator. And one of the things that we worry about is that potentially people at different income levels might have different digital footprints and different availability of that kind of data. So, you think about credit cards scores; someone who doesn't have a credit card might not show up.

And I think you had this data on 70 percent or something of the population. So that's just a caution, I think, as we move into this world of trying to creative with measures. It's something to be aware of.

And I also wanted to note I'm a little confused about the Medicaid part because in the second measure with the population-based total cost of care you have the analysis showing that when you put Medicaid into the model that there was more utilization from a cost perspective by $75 a person. And so but the cost of Medicaid is reimbursed so much less that it averages out to be less.

So, that's just something to keep in mind. I know Health Partners' recommendation is that these are used within populations so you would use this only for Medicaid, only for commercial. But it's just something to be aware of, that that is potentially an indicator of the impact of sociodemographic status on these measures.

And then just to share a little bit about some of the research we are doing at my organization. We've been looking at the impact of homelessness on different kinds of measures, quality measures, utilization, et cetera. And one of the things that we're learning is that for something like length of stay that can be kind of a bimodal distribution, so you can have people, for example, experiencing homelessness might have, there's a group of them who might have lower length of stay than typical for a certain DRG. And we think that's partly because those folks are about the kind of social supports available. But medical diagnosis might actually

require hospitalization in order to keep them safe because they don't have those community supports.

And at the other end of the spectrum we have people who are in our hospitals for 6, 9 months because there's nowhere else for them to go but they're medically stable. And so the truncation of the outliers is going to wipe out some of that, those effects that you see. But it also might, the other side of it might be pulling that effort down.

And so there may be reasons conceptually that we're not seeing a relationship. But I think we'll be learning more about it, so.

MS. KNUDSON: So, Nancy, thanks for those comments. It's important. And you got the results. You also understand that it's very important to understand the population that a provider might have as a whole, but then also appreciate how we're recommending segmenting these measures so that you can really get at that

action and validity of pieces.

So, for this, for this commercial population, you know, it sounds like you, you have congruent comments with our findings. But as any of us serving a homeless population and also a large Medicaid population, it's important for us to understand different methods to measure and inform appropriate interventions for those population.

But this is a commercial measure. As Gary described, the clinical risk adjustment, as well as the fact that it's segmented to a commercial product, really drove these results.

CO-CHAIR ASPLIN: Thank you. Janis.

MEMBER ORLOWSKI: I just want to go back to my comments and state I find it -- I, you know, and a number of other people here may know and may be familiar with the commercial risk adjustment methodology. But I think it's inadequate to use it for a measure that there's not full transparency in that measure. And to say that we have utilized this opaque methodology

and we've looked it and we find no difference and, therefore, there's no disparity, it's just an inadequate measure for an NQF standard.

And I think what we have to do is we have to be able to fully look at all the attributes, you know, of disparity measurement.

CO-CHAIR ASPLIN: Kristine.

MEMBER MARTIN ANDERSON: This is a little bit in a different direction, so I just want to put it out there.

When I think about validity of these types of measures that are used in partnership with other quality measures to tell something about value, to me the validity test that you really want to have is when you make a change and then you impact this resource use measure, that you are then able to show that the quality at least maintains the same -- you know, there was an impact; right? So you make a change and there was an impact on the measure. That's test one.

And then test two is that then when you look at the quality measurement you are

seeing the right interaction.

And so I, I do think that as we think about resource use measures and the validity of them, for what they're meant for in the over all purpose you need to figure out how to link it to the quality measures and, also, how to link it to intervention so that you know whether or not someone can affect it; right?

So, maybe the easiest ones are things like correlations with readmissions or things that you know are going to relate to resource use. But the harder ones I think around the other quality measures.

Because I think when we say, oh well, cost and quality are not correlated, we're saying that largely because they're both all over the board; right? And they can be in multiple directions. I just think there needs to be more work to try to figure out at least the value and the validity question.

CO-CHAIR ASPLIN: Thank you. Cheryl, Jack, and then Nancy you have another comment?

CO-CHAIR DAMBERG: Oh, I just wanted to follow up on Kristine's comment.

So, I view validity as do providers have control over their ability to influence things, recognizing multiple factors will influence this outcome? And I think the issue of SES comes into play in terms of their ability to affect all of the utilization of care that may be also under control.

So I think that's sort of the issue that's on the table.

And, you know, I personally have some concerns about the extent to which they fully explored the SES space here. And so I do have reservations on that front.

CO-CHAIR ASPLIN: Jack.

MEMBER NEEDLEMAN: I want to make a comment on the SDS methodology and then to a couple of other aspects of validity.

I don't fault the developers here on their SDS testing. We have seen this kind of approach done before on other measures and it

seems to be the way people have been presenting the data. But I think -- and this is the point -- when we are concerned about SDS, to some extent we're concerned about unmeasured need for populations that are disadvantaged that are not captured in the risk adjuster. And for that, the kind of regression analysis that was done here may make sense.

But there are at least two other aspects at the provider level that we're concerned about SDS. And this particular methodology does not speak to it. And I think NQF has to think about how to deal with these issues.

One is the issue that providers that disproportionately treat the poor and disadvantaged populations have fewer other resources to deal -- to provide care to those folks. And, therefore, it shows up in increased use of other resources on things like readmissions.

And that's a provider-level data. It's based upon the proportion of the population

in the practice that is disadvantaged. And the individual level measurement doesn't get at it.

And the other issue that we're dealing with is the issue Nancy raised, same thing, higher proportion of patients who are disadvantaged in communities with fewer resources, within fewer resources at family level, and the provider has to replace those with additional care. And, again, that's going to be something we see at the group level based upon the aggregates in the care that's provided for those populations. It's not going to show up, clearly, in a linear regression in which the high income people and their experience complete offset the lower income people in a way that does not capture the concern that we have about SDS at the provider level.

So I think NQF really needs to think about what guidance you need to, we need to give people about these measures if we're really going to understand the impact of concentrated low income or disadvantaged populations in provider

groups and their impact on the care that those groups are providing. So, I want to say that.

In terms of the other issues of validity here, I think we come back to the issue of how much variability can a group expect in this measure just from the luck of the draw, not because of poor performance or anything. This is a measure with very tight variance in that 10 to 90 percent population. So, if a 1 percent, 2 percent variation year by year, not about reliability but just the luck of the draw of who the patients were this year, is there then it could completely substantially shift where a group is ranked but also what the group ought to do. If they're reacting to small, you know, random variations in use around individual patients then I think you've got a problem here.

And I think we need to think about the variability, within-group variability from random years, not the patients, random years versus the degree of variability we're seeing across the groups. And I'm not sure it's that high. I

think it may be a problem around the resource use measure here, but I'm not quite sure about it yet.

MEMBER WEINTRAUB: So this is Bill. Let me comment. I tried raising my hand. I don't find it's working very well.

Because I think Jack raises a very important point. But, you know, this runs through, this problem of stochastic error runs throughout when we're trying to apply broad -- apply measures broadly to large groups. Some people are just going to stand out just on a stochastic basis not looking very good. Or others may look particularly good when they're not good. And that's in the nature of this.

People have tried to use statistical approaches like Bayesian hierarchical modeling. And I don't see that within this measure.

Have the developers thought about that?

MS. KNUDSON: So this is Sue. I'm going to give a shot here with a high, kind of

high level response, and then would ask Gary and Chad to comment as well on some of the details.

You know, this measure is not going to solve all of our problems from a population health issue when we think about the full panel of patients we care for across the full spectrum of commercial, Medicaid, and Medicare. When you wrap in the different ages that people are and how that might customize the needs that they have and/or the interplay with their income levels and education and health literacy.

And, honestly, even for those of us who are well-educated, sometimes when you are thrown in the middle of a catastrophic diagnosis you are not at your best performance level. So, we can all -- if I put my care delivery hat on -- we can all relate to the things you are saying.

As it relates to provider accountability, I think it would go back to the culture question I would have. You know, that comment I made earlier. And moving, seeing directional improvement in both the resource use

metric and quality, that we have examples that we've worked with a provider in Northern Minnesota, for example, that the resource use, really high in the pharmacy component for their patients with diabetes. They realized they had some large group of patients on the higher end of BMI, and they were prescribing pens, which were very expensive. Patients were going to multiple.

They switched over to vials. And, you know, it was both a price piece, so getting into total cost of care, but a resource use component that sort of led them there as well. Just and then you sort of keep digging. And at the same time realized better quality improvement in their quality results for those patients.

So, the examples are there. It goes to accountability. We're just looking for directional movement at the same time.

And then getting back to the SES component and what we can do as groups, I think it's important. I know as a provider ourselves -- and I'm not a clinician but I work side by side

with the clinicians, and I in my career have been an administrator in our care group. And our approach is really to understand individual patient needs.

At the same time we're understanding our overall population health results. And that means segmenting the population, as we have talked about, in some important ways, but also tailoring those interventions, whether it be for low income people, we will understand who's at higher risk for a readmission and wrap some community services around them. Whether it's our partnering with the firefighters for post-discharge visits and really identifying food scarcity issues that are the results for the readmission.

But that's a strategy largely in our low income population. So I'm just saying it's sort of an all hands on deck.

And for our commercial population for which these measures are intended to endorse -- be endorsed, we have, we have a purchaser

population, large, small, and independent groups that are funding their group insurance. And we also have that important group as a constituency to make sure that we're providing high quality, most affordable solutions for them.

And so it's, you know, as a provider we've got to be accountable across the whole. And as a payer, we also need to help providers by segmenting these measures. And what we have in front of us today is sort of a commercial, that commercial segment.

So, I know it gets very challenging when we think about the broad panel. So those would be my sort of high-level comments just acknowledging the comments and concerns that are being raised, but also trying to hone in on what this measure is intended to do.

CO-CHAIR ASPLIN: Thank you, Sue.

Srinivas.

MEMBER SRIDHARA: Yes. Just a comment on the SDS work. I haven't been engaged with NQF's work in it in the past. So just as a -- I

may be missing the point perhaps; I don't know.

So, from a public health background, critical obviously, and we need to measure it. However, I think in practical terms, if we skip ahead a few sections to usability or feasibility, if you insisted on being able to have information that could tie particular location level of this kind of information, this would be nearly impossible to implement anywhere for anything. It's a -- at least if you're going on a claims-based analysis.

So unless you happen to have an approach to do that, I'm not sure what else they can do other than these sort of applications of, you know, census tract or other data that you might get from census, or if you take the experience-type data that they have. I'm not sure you can realistically implement it. I think it's a worthwhile line of questioning how you might use it or as guidance to the end user. But I certainly wouldn't want it to be a part of the measure definition in any way because then it

would make it nearly impossible to use because most of them would not have access to that kind of information.

So just a comment.

CO-CHAIR ASPLIN: One quick comment I'll make on validity as well. It's somewhat usability but it speaks to the validity.

So the context of these measures, and I'm going to speak now from the standpoint of leader at Fairview which had a shared savings contract with HealthPartners, had a commercial population at risk. It's always in the context of this, paired with the total cost measure, paired with your contract performance. Right?

So you have a shared savings agreement. And I'm not saying this solves all the validity questions, it's just I never had a situation where the three didn't match. Right? So, you know, you're upside down on your shared savings but your resource use looks really good; never happens. Right? So they were always in agreement relative to the results of our shared

savings outcome. And then these two measures paired really helped us try to understand where practice patterns may be driving cost and research use versus the pricing problems were driving total cost of care challenges.

And so from that standpoint it all held together very well in terms of validity. And, you know, I found it very useful from that perspective.

Helen.

DR. BURSTIN: Thanks. I just want to make a brief comment to some of the broader conceptual issues raised by Jack and Janis about the over all approach.

One thing we have heard repeatedly is this is continuing to be a difficult issue. At times I think it's fair to say we don't -- the data are not yet available in a form that's readily accessible to perhaps get at the best variables. That's number one. And we continue to struggle with that throughout the process.

I think the second issue is that the

original work of the SES panel -- and some of you were on that original panel -- specifically focused in on individual-level variables, not provider-level variables, so not percent Medicaid, for example, but individual level factors.

And, actually, one of the things after a recent discussion on recent readmission measures we looked at is our Disparities Committee in the next couple of months is going to spend a section of time specifically thinking about how would we begin thinking about the use of the provider-lever and community-level factors. Because, again, it's something that keeps coming up. A lot of concerns were raised that that may not be the best approach in terms of adjusting for provider-level differences and then being able to see provider-level differences in the outcome.

But those are issues very much on our radar. I think, you know, this particular measure at least laid out their methodology. And

I think you have to judge it based on the data they have available at this time and whether you think it's adequate certainly.

And then lastly, at least for the measures we have been looking at to date, including the recent cost measures as well as the readmission measures, the first issue I mentioned about the availability of adequate data to assess SES and SDS has made us indicate that these measures should come -- be reexamined on an annual basis as part of the annual update for the measure developer to ask about the availability of newer adjusters. Because I think as we get better adjusters -- and again, we're going to continuously be left with this question of are the adjusters adequate to the past of what's available? And if we had better adjusters we would find more differences. And then as we begin to explore, how do we move into looking at provider and community-level factors which other papers in other areas have found to be quite significant?

CO-CHAIR ASPLIN: Janis.

MEMBER ORLOWSKI: Just to respond. So, Helen, that sounds great. And back to what Jack was saying.

I think when you take a look at this the question is is if, for example, in a community there is 20 percent of the population that has increased needs, whatever the increased needs are: whether they're dual-eligible; whether there's high mental health issues; whether there's, you know, poverty; what you want to do is you want to be able to incent the providers and the hospitals and, you know, everyone within the health field to continue to provide care to those individuals and not e "hurt" in a measure like this.

So if 20 percent of your population ends up to have higher burden of care, what can we do so that, you know, there's, there's two responses: one is, you know, we're just going to be on the lower end of this measures. That's response one.

Response two is, you know, you -- I try to limit the number of individuals that you care for. Completely the wrong thing that you want to incent.

And so, I like the idea of taking a look at the burden of, you know, sort of the burden of the SDS within the community. I think that's a very interesting and probably gets us away from this incremental, you know, patient by patient, you know, what are the factors that address them.

So, very nice.

CO-CHAIR ASPLIN: Very good.

MEMBER WEINTRAUB: Good in principle, but can you do it? Can you get the adequate measures to allow you to accomplish that? That's always the question.

DR. BURSTIN: At least on the readmission side, for example, people have been pulling some community-level indicators. For example, housing vacancy in a health affairs paper last year off census looked to be

important. Again, those are probably indicators of what's happening.

The other piece I just want to mention to this group as well is we are increasing recognizing, even before the impact report, but particularly since the impact report came out just a few months ago on SES factors, that some of what we think may be SDS is probably unmeasured clinical complexity. And if you look at the SDS -- if you look at the impact report, for example, they raise issues of, you know, measures of frailty are probably playing into percent dual-eligibility influence and at least on the Medicare side.

So I think these are going to be issues we're going to try to tease out. I think it's probably just time for all of us to have sort of a risk adjustment 2.0 reboot. How do we really look at risk adjustment broadly? And how much are we picking up that's SDS that's really clinical complexity that we need to recognize and take care of, as opposed to some of these other

factors. So I think it's a really big, important issues.

Just one last quick thing. The other thing the SES panel had said when that report came out a couple of years ago is risk adjustment is just one approach. And I don't want to forget that. Payment differences, as we just saw, is part of the 21st Century Cures Act in terms of comparing. Like providers to like providers is another.

The third one that this group has emphasized was stratification, you know, the fact that they have emphasized the use of stratification by payer is something you could include, for example, as part of the measure going forward.

CO-CHAIR ASPLIN: All right. So I'm just doing a -- going to pull back for a second -- time check. We lost our break. We were supposed to be about done with 1604 at this point. We already are going to have a working lunch. So just, just saying. Try to keep your

comments concise.

Nancy, Mary Ann, then we're going to vote on validity.

MEMBER GARRETT: So just one thing to add to Helen's list of possible alternative approaches to risk adjustment. Another one that was suggested by a recent National Academy of Sciences report is actually identifying measures and labeling them as to whether they are sensitive or not to social disparities. So I mean like a diabetes measure where there's lots of outcomes in the measure, like is your blood pressure in control, et cetera, highly influenced by the resources you have available to you. And something like central line infections in the hospital maybe not so much.

And if we can categorize measures according to how sensitive they are, then that would inform how they're used to move money around between, you know, systems and that kind of thing. So just another thing to add to your idea.

DR. BURSTIN: And that's actually part of our disparities work as well. We're going to take up our old disparities sensitivity algorithm and refresh it with the current Disparities Committee and think about -- and the impact report also very clearly said you need equity measurement. And we think that's part of it. We see this as really an interconnected set of issues. And this is just bringing it around.

CO-CHAIR ASPLIN: Mary Ann.

MEMBER GARRETT: If I could make one other quick comment.

CO-CHAIR ASPLIN: Oh, I'm sorry.

MEMBER GARRETT: Sorry.

Just I, I just want to also make the point that while we're talking about a commercial population for this measure, there's certainly a lot of disparities within commercial populations. And just as one example, my health system runs a small food shelf. And a lot of the patients that we give food to have commercial insurance.

So I don't want us to assume that

there aren't issues within commercial populations as well as we're thinking of this.

CO-CHAIR ASPLIN: Mary Ann.

MEMBER CLARK: Yes, hi. Just a couple of things.

One is the discussion about -- I don't know if we've discussed the pharmacy costs and how to account for that. Before it was estimated for the patient members who did not have -- who had it paid for through a PBM, for example. And I'm trying, I'm trying to remember that percent of members where that was the case.

Can you remind me of that?

And then also, if it was a large percentage was there any type of sensitivity analysis on that? You know, I know that you, you used an average for that, for that provider, I believe, on pharmacy costs for the patients who did not have pharmacy cost data.

MR. HEIM: Yes, this is Chad.

Our carve out rate for data that we're doing the testing on was about 20 percent

pharmacy carved out.

MR. KITCHING: And then -- this is Gary -- and then that will vary depending on the provider because, obviously, providers are going to have different rates of carve-outs. And tested that down and the lowest one we got to was about 70 percent of people had carve-outs -- or had pharmacy at a 30 percent carve-out rate.

We're going to do an analysis right now that kind of tests that all the way down to see where it be breaks, to be honest. We never did that before because we never experienced it in our market. Most folks, to be honest, have a high -- only a low percentage of that carve-out. And that came back invalid.

So we're going to try to test that down to, again, see where that carve-out level breaks. And we'll have it on our website at some point in the future.

MS. KNUDSON: But we haven't had that be an issue with the users so far.

MR. KITCHING: Yes, it hasn't been an

issue in our market or anybody across the nation that have actually depended on the measure. Again, just for because we want to do our due diligence, we want to make sure we test that down.

MEMBER CLARK: Okay, great. And then just one other questions I had.

We haven't talked about validity. And I know -- I mean face validity. And I realize that, you know, the important thing is the empirical testing. But I know that you did some extensive face validity testing as well. And that's kind of where, in one sense where, you know, it's pretty valuable to your, you know, your users there.

So I was just curious if you might be able to summarize briefly what was found in that face validity testing?

MR. HEIM: Yes. So, a couple examples, one with our business. When we run our, run our total cost of care and resource use results we make those available to our provider

networks for them to review and vet. And we have a 45-day comment period for them to review those results.

And in addition, we meet frequently internally with medical directors in a broader relation in going through those results. And as well as then meet with the providers frequently to walk them through those results.

Other examples for face validity is implemented at measurement where there's a multi-stakeholder committee with numerous stakeholders having, in essence going through NQF endorsement through over a year-long process re-vetting the measure and providing input on that as that was rolling out.

And then same was true when we rolled out research use with measurement.

And then the other example would be NHRI who have -- are just like measurement, they have regional stakeholders involved reviewing and vetting the measures as well.

MEMBER CLARK: And in terms of what

you found though, in general there were no issues? Or if there were issues, how were those, you know, addressed?

MS. KNUDSON: You know what, we did not have issues with the measure. I think most of the things that we -- if we ran into anything it was in the "how to" in rolling this out. But it wasn't in the measure itself.

And so we haven't had any of those issues. And I think many of the national organizations who have taken up the measure, or our local constituents who wrote in positive comments would also help with that face validity piece, as well as Brent's previous comment. So, thanks for that from usability from a Fairview perspective.

That's commonly what we hear from organizations. And, you know, we're so super transparent because of the idea that we want to compete on our results, that I think that really helps particularly our clinical colleagues because they've never really seen anyone be this

transparent with methods. And all that consultation and data sharing to drive an improvement portfolio that goes along with it.

So it's partly in the "how to."

MEMBER CLARK: Thank you.

CO-CHAIR ASPLIN: All right. Thank you.

Let's move to the vote on validity. We've had a good conversation, good discussion. And unless I see any objections to moving to the vote, I think we should move forward with that. So, Irvin.

MR. SINGH: Thank you. So, we're going to begin voting on validity for Measure 1598. Your options are 1, high; 2, moderate; 3, low; or 4, insufficient.

Please vote now.

(Voting.)

MR. SINGH: So all votes are in. Voting is now closed for validity for Measure Number 1598. The results is as follows:

2 votes for high; 14 votes for

moderate; 2 votes for low; and 0 votes for insufficient. That accounts for 11 percent of the votes being high; 78 percent of the votes being moderate; 11 percent of the votes being low; and 0 percent for the votes for insufficient.

The measure passes for validity.

CO-CHAIR ASPLIN: Thank you. Bill, you have -- I think just from a time check standpoint we're going to try to plow through and finish 1598. And then we'll break to grab lunch for about 15 minutes and then do a working lunch and consider 1604. So that's the overall game plan.

And, Bill, to your comments, please, as the discussant on feasibility.

MEMBER WEINTRAUB: Right. So, obviously, this is a much simpler one than what we just discussed.

The measure has been around and it's been retested. It's been shown that people can apply it. However, so we seem to be really

feasible from that standpoint. Also, no new data is required; it's all administrative data.

Again, it suggests high feasibility. There are a couple of caveats though. Well, I mean this really applies to the commercial space rather than Medicare and Medicaid, which is fine as long as we keep within that space.

The other caveats are that the data we're seeing are from Wisconsin and Minnesota, as I understand it. This is applied widely. But we're not seeing data quite as widely as we might.

And, finally, that it does require applying some commercial software. So there is some expense related to it.

But that's really it. I mean, I think that this is really pretty simple and does not require additional data collection, which is very good.

CO-CHAIR ASPLIN: Thank you, Bill.

Any other comments or questions relative to the feasibility of Measure 1598? I

see none in the room and I do not see any hands raised online.

So, without objection, I think we should move forward to the vote on feasibility for 1598.

MR. SINGH: Okay, thank you.

So we're going to be voting on feasibility for Measure Number 1598. Your options are as follows:

Number 1, high; number 2, moderate; number 3, low; number 4, insufficient. Please begin voting.

(Voting.)

MR. SINGH: So all votes are in. And voting is now closed for feasibility for Measure Number 1598. The results is as follows:

14 votes for high; 4 votes for moderate; 0 votes for low; and 0 votes for insufficient. That accounts for 78 percent of the votes for high; 22 percent of the votes for moderate; 0 percent votes for low; and 0 percent votes for insufficient.

This measure passes for feasibility.

CO-CHAIR ASPLIN: Thank you, Irvin.

Jack, you're the lead discussant for usability on this measure. Your comments, please.

MEMBER NEEDLEMAN: Okay. So, again, I think the conversation here can be relatively short. The first measure of usability is it's used. And it has been wide -- it has been extensively used. If anything, according to the documentation we received, the use has expanded, which suggests a high level of usability.

When I look at the measure and when I look at the comments that have been made two or three things stand out in terms of usability. One is part of what makes this usable is the disaggregation by type of service that is provided. So you get those as well.

And without that it would be a lot more difficult to make use of the measure. But that is present. That seems to be part of the methodology, so that solves that problem.

The attribution issue was raised in some of the online comments about the appropriate, you know, usability. Similarly, comparison groups were raised in some of the online comments, with specifically the example of gynecologists being treated as primary care providers and yet having potentially very different kind of resource use pattern than other primary care providers. That would seem to be solvable in terms of the groups that are using it, in terms of getting appropriate comparisons.

But overall, the usability looks like it meets those criteria by virtue of the fact that it's been widely used. And folks like Brent have indicated that it was useful to them.

CO-CHAIR ASPLIN: Thank you, Jack.

Any comments, additional comments before we move forward?

Yes, I would just for my own perspective add quickly, I had opportunity to work with ACOs, track 1, track 3, primary ACO, commercial shared savings, Medicare Advantage,

with, in multiple markets with regional carriers and national carriers. And, quite frankly, for feedback on cost and resource use, the HealthPartners measures are completely in a league of their own. There's nobody that's close, in my experience, in terms of actionable feedback on your cost and resource use.

Nancy?

MEMBER GARRETT: So, I was just going to share that my experience as a provider in the market in Minnesota where Minnesota Community Measurement has adopted this measure, and I think it's great that HealthPartners is providing a lot of the kind of analytic tools and ability to disaggregate the data.

Community Measurement, so what they're doing is they're calculating this on a system basis. So, for my medical system I get a report for commercial patients showing these two measures, the resources and the total cost of care across all my commercial patients. And it's super helpful because otherwise we're getting a

different report from the rate payer, and they all say different things. And so now this aggregates together in a much bigger population.

But that organization isn't adopting all of the tools that HealthPartners provided since we're in a contract with them, and so in that sense it's been less usable because keep in mind that this measure, it includes utilization within your own provider system but also external utilization. So anyone attributed to you for primary care, it's all the services they're using, no matter where they are. And if you don't have any information on that, there's, you know, a lot less ability to kind of use those figures.

So, that's not a fault of the measure but it's just kind of a practical reality of implementation of this.

CO-CHAIR ASPLIN: Thank you, Nancy.

I think we'll move forward, unless there's objection, to a vote on usability. Irvin.

MR. SINGH: Thank you. So, we're going to begin voting on usability and use for Measure Number 1598. Your options is as follows:

Number 1, high; number two, moderate; number 3, low; number 4, insufficient. Please begin voting now.

(Voting.)

MR. SINGH: So all votes are in. And voting is now closed for usability and use for Measure Number 1598. The results is as follows:

13 votes for high; 5 votes for moderate; 0 votes for low; and 0 votes for insufficient. That accounts for 72 percent of the votes for high; and 28 percent of the votes for moderate; 0 percent votes for low; and 0 percent of votes for insufficient.

This measure passes for usability and use.

CO-CHAIR ASPLIN: Thank you. So the measure has passed all the must-pass criteria, as well as feasibility and usability. So our final vote on this measure is the overall suitability

for endorsement.

And, Irvin, could you walk us through this?

MR. SINGH: Yes. So we're going to begin voting on overall suitability for endorsement for Measure Number 1598. And the options are number 1 for yes, and number 2 for no.

Please begin voting now.

(Voting.)

MR. SINGH: John, if you could please submit your vote now.

MEMBER RATLIFF: Excuse me. Sorry for the delay.

CO-CHAIR ASPLIN: So everybody's voted. We're at 17. We've been at 18 the last few.

MR. SINGH: John, can you please email your vote.

CO-CHAIR ASPLIN: While we're waiting for that vote, Operator, if you could prepare. Our published agenda has opportunity for public

and member comment right now on the published agenda. And so, before we break for lunch we just want to honor that published agenda in case someone has dialed in and would like to make a comment at this time.

Irvin's got the results here. And then, Operator, if you could open up the lines for public comment.

MR. SINGH: So all votes are in. And voting is now closed for overall suitability for endorsement Number 1598. The results is as follows:

18 votes for yes and 0 votes for no. It accounts for 100 percent for yes.

CO-CHAIR ASPLIN: Thank you. Thank you to the committee. I get an F for honoring breaks so far today. We'll try to improve, although we're behind the 8 ball now. But, hopefully, we got a lot of the comments out. And 1604 we'll have some similar topics I'm sure.

But, Operator, could you see if there are any public comments at this time.

OPERATOR: Yes, sir.

At this time if you would like to make a comment, please press star then the number 1.

Okay, and at this time there are no public comments.

CO-CHAIR ASPLIN: Thank you. And thank you to the HealthPartners team. We're going to break for 15 minutes and reconvene. The lunch is available here in the room. So 12:15 Eastern we'll start with the conversation on 1604. Thank you.

Sorry. Are there any comments in the room on 1598?

(No response.)

CO-CHAIR ASPLIN: All right, very good.

(Whereupon, the above-entitled matter went off the record at 12:00 p.m. and resumed at 12:18 p.m.)

CO-CHAIR ASPLIN: Okay. We are going to move ahead and get started with 1604. And I've been informed we have an option on

importance to measure.

We have an option on importance to measure for Measure 1604, given the similarities between Measure 1604 and 1598. That's 1604 and 1598. The Committee has an option to roll its votes from 1598 and having them apply on 1604 just for this category, for importance to measure.

Unless -- if there's objection to that, we can hold a separate vote. And if there are specific comments on importance to measure that people would like to make in distinction for 1604 versus 1598, we obviously will entertain those.

Larry, you were the lead discussant, did you have any differences in terms of your comments on importance to measure for 1604?

MEMBER BECKER: Not substantive.

CO-CHAIR ASPLIN: Is there objection to using our votes for importance to measure only from 1598 and applying those to 1604?

(No response)

CO-CHAIR ASPLIN: Seeing none, anyone online have an objection?

(No response)

CO-CHAIR ASPLIN: Hearing none, we will apply the votes for importance to measure from 1598. And it will pass that criterion and that includes all three votes, correct?

MS. DUDHWALA: All three votes.

CO-CHAIR ASPLIN: Very good. So we will move to scientific acceptability. And Mary Ann, I apologize because I misread the notes that were very clear now that I read them. And when I called on you last time.

But, Mary Ann is our lead discussant for scientific acceptability for 1604.

Mary Ann are you online? She maybe thought we were going to take a little bit longer on importance to measure.

Mary Ann, if your --

MEMBER CLARK: Sorry about that. I had a little emergency I had to attend to here. But I am back now. So, we're talking about the

next measure.

CO-CHAIR ASPLIN: Correct. We applied our votes for importance to measure from 1598 to 1604. So we went through that with lightning speed.

So now we are on scientific acceptability. And we do have to do a full discussion of the acceptability.

MEMBER CLARK: Okay. So I need to vote on the other one I guess, right?

CO-CHAIR ASPLIN: You do not. We --

MEMBER CLARK: Oh. Okay.

CO-CHAIR ASPLIN: Didn't do individual votes. So we accepted their votes from 1598 just for the importance to measure criteria. Those three votes.

MEMBER CLARK: Okay.

CO-CHAIR ASPLIN: So, take us through the scientific acceptability on 1604 if you would, please.

MEMBER CLARK: Sure. Let me just bring up my documents here. I think in general

it's going to be fairly similar, you know, to the resource use measure.

I mean, but we're looking at it from a cross perspective where a, you know, allowable -- allowed payment perspective. So, that's really the difference.

You know, the same -- the data has been updated, you know, from the previous endorsement. So we have a new data set being used.

But basically the numerator is the total per member per month medical costs, and pharmacy costs. And then of course risk adjusted with the same ACG risk adjustor. And then that being divided by the comparison group measure.

In terms of the exclusions, it's the same exclusions as before. Members under the age of one, members 65 and over, and members with less than nine months of enrollment.

Let's see, commercial payer claims. And for the socioeconomic testing of course the Medicaid data was used as well.

And let's see, the measure was tested at both the data element and the score level on the -- for reliability. And again, using the same methods as in the resource use measure, the 90 percent random sample and bootstrapping sample as well as changes over time.

Let's see -- the results, you know, of those two tests for reliability, they were basically what we talked about for the other measure. So, I don't know how much we want to go into that here.

I mean, there was some data that looked at, you know, the results of the first endorsement compared to this endorsement, I guess. And they did change slightly, I guess.

It looks like the variance actually got a little smaller. And let's see, it looks like in terms of the provider performance across the years that the data was relatively consistent across all three years with a difference on average of about three percent.

So, let's see, what were some of the

issues? I think in general people felt like this measure was reliable.

And I think in terms of the concerns voiced, they were very similar to the resource use measure. Had to do with attribution methods. Some people wanted to know more about the risk adjustment method being used.

But, in general, I think, they -- the comments were that the reliability was high and that the measure is clearly defined and implemented.

Let's see here, the same -- a similar concern is on the other measure about the testing of the data that was in a one payer that was restricted to, you know, a certain geographic region of the country. And whether that would have an impact on the generalizability of the results.

So, I think, -- let's see. In terms of answering the questions for reliability that were posed, let me get to those here.

Are all the data elements clearly

defined and the appropriate codes included? Is the logic in the calculation clear? Can it be consistently implemented?

And is the construction logic clear? I don't think that most people had issues with that.

In terms of the testing, is the sample adequate and generalized to widespread implementation? And do the results demonstrate sufficient reliability so that the differences in performance can be identified? Yes, so again, these testing results for the empirical test were fairly similar to the resource use results as well.

So, I don't know if anyone has any other questions or anything to add to the reliability.

CO-CHAIR ASPLIN: Thank you, Mary Ann.

MEMBER CLARK: Yes.

CO-CHAIR ASPLIN: I appreciate it. So, let's start with reliability in terms of comments, questions, concerns about reliability

on measure 1604.

I do not see any hands in the room or online. Nancy, do you have any comments on reliability?

MEMBER GARRETT: No.

CO0CHAIR ASPLIN: All right. Very good. Let's move forward to the vote then for reliability on 1604. Irvin?

MR. SINGH: All right, thank you. So we're going to begin voting on reliability for Measure Number 1604. Your options are one high, two moderate, three low, or four insufficient. Voting has now begun.

(Voting)

MS. DUDHWALA: Andrea Gelzer, we're just voting on reliability for Measure 1604. If you could enter your vote.

MEMBER GELZER: Did you get it?

MS. DUDHWALA: Yes. We got it.

MR. SINGH: Okay. So all votes are in. And voting is now closed for reliability for Measure Number 1604.

MS. O'ROURKE: We got an extra vote. So, to avoid any corruption concerns, if we could all just vote again.

MR. SINGH: All right. So we're going to begin voting again on reliability for Number 1604. Please vote now.

(Voting)

MS. O'ROURKE: I think we're missing two. So if you could -- if everyone could hit their clicker one more time.

MEMBER WEINTRAUB: Did you get the phone-in votes?

MS. O'ROURKE: I believe we have all the phone-in votes. So, the missing vote -- oh, we have 18.

MR. SINGH: All right. So all votes are in and voting is now closed for reliability for Measure Number 1604.

The result is as follows: ten votes for high, eight votes for moderate, zero votes for low, and zero votes for insufficient. That accounts for 56 percent of the votes for high, 44

percent of the votes for moderate, zero percent of the votes for low, and zero percent of the votes for insufficient.

This measure passes for reliability.

CO-CHAIR ASPLIN: Thank you. We'll now open up the conversation on the topic of validity. Comments from committee members related to validity for Measure 1604. Jack?

MEMBER NEEDLEMAN: It's just, I need some clarifications from the developers of how price is paid or included in here. Again, around hospitals for example, some payers maybe paying DRGs. Others may be paying discounted charges.

Is the -- what is the basis for including -- for the prices that are in the model? Where there are variations in how the units of payments?

MR. KITCHING: So the full cost measure again, is using a loud amount, which the plan liability plus the member liability, and then looking at -- specifically at the riders. The prices select, for example, hospitals, could

be varied. They could be on a DRG payment system or they could be on a discount or fee-for-service basis as well.

But all those payments are aggregated up from the hospital from what the plan is paying, but as well as the member liability. And that's rolled together for the total cost of care for the all-in-one.

I would just add that the majority is prospectively priced.

CO-CHAIR ASPLIN: Thank you. Other questions from the committee relative to validity?

(No response)

CO-CHAIR ASPLIN: Seeing none online and none in the room, let's move forward to the vote on validity for Measure 1604. Irvin?

MR. SINGH: Thank you. So, we're going to begin voting on validity for Measure Number 1604. Your option is as follows: number one high, number two moderate, number three low, and number four insufficient.

Please begin voting.

(Voting)

MS. O'ROURKE: So we have another phantom vote. It got to 19. So we're going to need to do it one more time in the room. The folks on the phone, we have your votes. No action needed on your part.

People in the room, please vote one more time.

(Voting)

MR. SINGH: All right, so all votes are in. And voting is now closed for validity for Measure Number 1604.

The result is as follows: three votes for high, 14 votes for moderate, one vote for low, and zero votes for insufficient. That accounts for 17 percent of the votes for high, 78 percent of the votes for moderate, 6 percent of the votes for low, and zero percent of the votes for insufficient.

This measure passes for validity.

CO-CHAIR ASPLIN: Thank you. We'll

now move onto feasibility on 1604. Bill, do you have any additional comments relative to feasibility for Measure 1604?

MEMBER WEINTRAUB: No. Exactly the same comments I had on the previous measure.

CO-CHAIR ASPLIN: Thank you. Now open for committee discussion, comments, questions related to feasibility.

(No response)

CO-CHAIR ASPLIN: Seeing none in the room, I see no hands raised online. Let's move forward to the vote on feasibility for Measure 1604. Irvin?

MR. SINGH: We're going to begin voting on feasibility for Measure Number 1604. Your options are one high, two moderate, three low, or four insufficient. Please start voting.

(Voting)

CO-CHAIR ASPLIN: In the room, just hit your button again in case we're missing one.

MR. SINGH: Okay. So all votes are in. And voting is now closed for feasibility for

Measure Number 1604.

The result is as follows: 12 votes for high, six votes for moderate, zero votes for low, and zero votes for insufficient. That accounts for 67 percent of the votes for high, 33 percent of the votes for moderate, zero percent of votes for low, and zero percent of votes for insufficient.

This measure passes for feasibility.

CO-CHAIR ASPLIN: Thank you. Relative to usability. Jack, any additional comments on 1604?

MEMBER NEEDLEMAN: Yes. I actually have a question for the developers on this. The earlier measure uses standardized prices to -- so it's basically utilization weighted -- standardized price weighted utilization measure.

And then you get that disaggregation by different types. This includes the actual prices paid. So, you see a price-weighted utilization measure, where actually the prices are being paid.

If in terms of usability you want to be able to know how -- you want to be able to disaggregate how much of the effect is associated with pricing, how much you're paying versus the volume of use.

And in the documentation, it's not part of the measure, but in the documentation there is a price index which looks like it's the ratio of the total TCI measure and the RUC measure.

I could not see in the materials provided whether that calculation is done down to the individual types of services that are being provided. Which would clearly affect the usability.

You have to know whether you're paying a lot more for hospitals then the average. Or whether you're paying a lot more for your E&M services then average in order to figure out whether you're getting value for those additional payments.

So, I'm just asking the developers,

does that calculation of the price index, not an inherent part of the measure, but part of your documentation, is that carried out down to the individual types of services there for which data is provided?

MR. HEIM: Hi Jack. This is Chad. Yes. And that's where the actual billing starts playing out when you're using these measures together.

But yes, it's a simple ratio of the cost divided by the research use. And then that's drillable all the way down to the procedure level.

So you can start understanding within a hospital setting or a professional setting, E&Ms if the price is driver or utilization is driving.

MEMBER NEEDLEMAN: Great. Thanks for the clarification.

CO-CHAIR ASPLIN: Other comments or questions related to usability? Troy, you have a comment?

MEMBER FIESINGER: Yes. A question for the developers. Can you ask them if the implementation cost as a pair of measures. I'm thinking of smaller systems or smaller physician groups that want to use this. It ties into my usability question.

MS. KNUDSON: You know, we don't really get into that at all with the users. It's -- but what I will say is most of our users have been able to bring this up with their existing staff.

And the only external cost is the commercial risk adjuster. But some of them, and many of them already have those licenses in place. So it wouldn't be incremental cost.

For some of the organizations that have brought it up from scratch like NHRI or Minnesota Community Measurement, the risk adjuster, at least the one Johns Hopkins makes available, they do have a much significantly lower cost research based license fee that is very affordable. Like around the thousand-dollar

range versus what us as commercial payers would pay.

So often these tools are already available by the users. And if they're not, they seem to be for these organizations that give a much more affordable price break and then existing staff.

And then we have seen some care groups, you know, work with their consultants. But they're not hiring like brand-new consultants. It's their existing relationship.

So -- but I don't have a specific number for you. But that's our observation.

MEMBER FIESINGER: Okay. That's still helpful. And then are you able to aggregate, it sounds like all the different commercial payers data, for the organization if you're contracted to it?

MS. KNUDSON: Yes. If you're contacted to do it as a vendor, we just require the vendors to, you know, the software is free of charge. But we do have a license agreement for

the resource use. And that is because we have a patent on our resource use.

And our whole idea was to put this stuff out there in the public domain free of charge. So we didn't want folks to commercialize it and then have additional funding be an impediment of users adopting it.

So we do require vendors to license it one client at a time. Which is super easy to do.

MEMBER FIESINGER: Okay. Thank you.

CO-CHAIR ASPLIN: Thank you. Seeing no other comments, let's move to the vote on usability for 1604. Irvin?

MR. SINGH: Okay. Thank you. So, we're going to be voting on usability and use for Measure Number 1604. Your options are one high, two moderate, three low, and four insufficient.

Please begin voting.

(Voting)

MR. SINGH: Okay. So all votes are in. And voting is now closed for usability and use for Measure Number 1604.

The result is as follows: 12 votes for high, six votes for moderate, zero votes for low, and zero votes for insufficient. And that accounts for 67 percent of the votes for high, 33 percent of the votes for moderate, zero percent of the votes for low, and zero percent of the votes for insufficient.

This measure passes for usability and use.

CO-CHAIR ASPLIN: Thank you, Irvin. And we can now do our final vote, overall suitability for endorsement. The yes/no question. We can move to that at this time, Irvin.

MR. SINGH: Thank you, Brent. So, we're going to begin voting on overall suitability for endorsement for Measure Number 1604. Your options are one yes, two no.

(Voting)

MR. SINGH: Okay. So all votes are in and voting is now closed for overall suitability for endorsement of Measure 1604.

The results is as follows: 18 votes for yes, and zero votes for no. And that is 100 percent of votes for yes.

CO-CHAIR ASPLIN: Thank you. So this concludes our discussion of Measure 1604. We are going to open up before the developers leave here, open up for public comment on -- I'm going to start it with 1604.

So, if there are comments online, Operator, can you open that up for 1604. And just for anyone who's online and has a comment for 2158, we're going to do that next before we start the conversation on it.

So, 1604 any public comments on the total cost index, 1604. Operator?

OPERATOR: Okay. At this time if you would like to make a comment, please press star then the number one.

(No response)

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

CO-CHAIR ASPLIN: Thank you. Are

there any comments in the room on 1604?

(No response)

CO-CHAIR ASPLIN: Thank you. And with that I'd like to thank Sue Knudson, Chad Heim, and Gary Kitching for your comments and support in terms of the conversation today. It was very helpful to have you on the line to answer questions.

So, thank you to the three of you.

MR. KITCHING: You're welcome.

MS. KNUDSON: Thank you.

CO-CHAIR ASPLIN: And with that, I'm going to hand this over to Cheryl. Because she's going to be leading the conversation on 2158. And we've had a request to do the public comment first.

So Cheryl, go ahead.

CO-CHAIR DAMBERG: Sure. So we're now focused on Measure 2158. The measure -- excuse me, the Medicare Spending Per Beneficiary.

This is a hospital only measure. My understanding is that there is a measure in

development that will focus on physicians.

But, before the committee begins its work, because I know folks on the line potentially have been waiting a long time, we're going to accept comments at this point.

So Operator, can you find out if there is anyone who would like to make a comment on 2158?

OPERATOR: Yes, ma'am. At this time if you would like to make a comment, please press star then the number one.

(No response)

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

CO-CHAIR DAMBERG: All right. And are there any comments from folks in the room?

MR. WETZEL: Yes. One comment. Hi, this is Scott Wetzel. And I am filling in for Janis Orlowski, who unfortunately had to leave.

But one comment we wanted to make was on SES and our concerns with that. The WUMC is concerned with the conclusions drawn from the SES

analysis of the Medicare Spending Per Beneficiary Measure and the impact that it has on hospitals who care for disadvantaged patient populations.

As highlighted earlier by Dr. Orlowski, we have concerns with the variables used. Particularly concerning the black/non-black variable. And request additional insights from the developer as to other variables that were considered for this measure.

We also have questions around the income-to-poverty ratio variable. Which resulted in a plus or minus difference of 0.01 or less for 97 percent of hospitals.

Due to the tight clustering of performance scores for this measure, a difference of 0.01 may significantly affect hospitals caring for disadvantaged patient populations that may require higher utilization of services. So, we request great detail as to whether any particular type of hospital was among those who were -- that moved 0.01 or more under this measure.

These comments were submitted to the

committee beforehand. So they are included on the worksheet as well. That's it.

CO-CHAIR DAMBERG: Okay. Thank you. Are there other comments before we get into discussing the measure?

(No response)

CO-CHAIR DAMBERG: All right. Seeing none. I am mindful of our time. Brent did a great job of getting us mostly back on track. But, we're running about 25 minutes behind schedule.

So, I would ask committee members to stay focused in your comments. And similar to the past two measures, I think most of the discussion is going to fall into the validity and reliability sections.

What I would like to do is begin by asking the measure developers to come to the table and give us a very brief overview of the measure. We have two folks with us today, John Pilotte from CMS and Sriniketh from Acumen.

So, please come to the table and join

us.

(Off mic comments)

CO-CHAIR DAMBERG: Okay. So for those on the phone, if you couldn't hear, Kim Spalding Bush is here from CMS in lieu of John.

DR. NAGAVARAPU: So, thanks very much for the opportunity to speak and answer your questions. I'll give just a very quick overview along the NQF evaluation criteria.

So the Hospital MSPB Measure measures total Medicare Parts A and B standardized allotted amounts during episodes of care for hospitals paid under the inpatient respective payment system across all conditions and admissions. The incentive care are designed to include the period immediately prior to, during, and in the 30 days after hospital discharge.

In terms of importance to measure and report, the measure addresses a high priority aspect of rising Medicare expenditures in a particular setting in which there are performance gaps.

The intent is to incentivize hospitals to coordinate care for patients and reduce unnecessary utilization and adverse outcomes during the period of time encompassing the inpatient stay and immediately after. And there are a large number of hospitals that are covered by the measure, over 3200 with at least 25 episodes and 20 percent.

The measure has been a well-accepted measure that served a purpose both on its own and informing hospitals about costs. As well as serving as kind of a tool to reorient hospitals towards the new emphasis in areas in CMS programs on bundled payment sorts of programs and care coordination.

In terms of scientific acceptability, in the pre-evaluation comments, we saw questions about the reliability threshold. And wanted to provide numbers for a higher reliability threshold of 0.7.

And 93 percent of providers at the case mean of 25, meet or exceed the reliability

threshold of 0.7. The requirement of a higher threshold then that maybe difficult given that there's natural variation spending across patients that we actually want the measure to reflect.

In terms of test/retest results from year to year, as was across random samples within the year, the Spearman ranked correlations and Pearson correlations are very high.

Especially considering if you think about kind of a time series model with just a standard moving average process, the maximum correlation you can get is 0.5. And so the correlation above a 0.8 in 2014 and 2015 is incredibly high.

In terms of validity, we know a lot of the discussion will focus on socioeconomic status. We did use income-to-poverty ratios at the zip code level as well as race defined by black and non-black.

We use those items in risk adjustment and then at calculated measure scores for

hospitals based on the new risk adjustment equations.

The results cited in the submission form showed that there's not much of a difference in terms of measure scores. With a large percentage of hospitals within a .01 change greater than 95 percent and 97 percent, depending on the specification.

The reason we use income-to-poverty at the zip code level is to strike a balance between the desire to look at individual beneficiary attributes, as well as looking at the community attributes that everyone's talking about in the previous session. Because we do think those sorts of community attributes are an important aspect of this that wouldn't necessarily be captured by a Census tract.

However, we saw the pre-evaluation comments. Realized that people wanted to see something that was more targeted towards individual beneficiaries. So we also ran an analysis recently looking at the fact of

individual dual status.

We ran risk adjustment specifications with dual eligibility dummies, recalculated the measures and looked at changes in the measure. For greater than 90 percent of providers, the change in the measure is less than .01. Sort of varying between .008 change on one end, versus .005 or 6 on the other.

The measures have -- and that's something we're happy to talk about. As we're familiar with the other analyses that we haven't fund along this dimension. And happy to talk about sort of the community attributes and how that factors into performance on the measure.

In terms of the other sort of standard metrics, for validity, the predicted ratios are remarkably stable across risk four deciles. And we can talk about the longer HCC look back and potential problems with that.

And the reasons why it doesn't increase predictive power to -- and it provides support for the 90-day look back.

The measure is highly correlated with spending at the hospital referral region level in an other publically available measure of resource use. And it's also correlated with service utilization rates.

We realize there are some questions and pre-evaluation comments about correlations with readmissions. And so we looked at that in order to see whether hospitals with readmissions tend to have higher or lower measure scores.

What we found is that for hospitals with less than 5 percent of their episodes having readmissions, the measure score was approximately .88 on average. For hospitals with greater than 20 percent of their episodes having readmissions, the hospital measure scores were greater than one with the hospitals with especially high readmission rates, up along 1.2 or so. So we wanted to be responsive to that.

In terms of feasibility and usability and use, the measure is reported in the hospital IQR, Hospital Compare, and Hospital Value-Based

Purchasing Programs. It's used for payment modification and goes through rulemaking each year.

And so there's rounds of public comments that are taken in every year. So the measure is well vetted and it's -- there's opportunities for hospitals to comment on the measure periodically, and for CMS to take into account, changes that are suggested by hospitals where appropriate.

And the measure is highly actionable. There's a question and pre-evaluation comments about how hospitals can make use of this. And how they can control their post-discharge spending.

The hospital-specific reports that come with the measure is as long -- as well as the episode files, allow hospitals to understand where spending is being driven by, which providers in the post-discharge period are driving most of the spending.

And that allows for physicians to

coordinate with providers in those specific facilities in the post-discharge setting. As well as to take into account other more systematic improvements across hospital coordination with post-acute care facilities.

The measure's been vetted through multiple periods of public comment. And we're happy to hear your feedback and answer your questions.

And so I can circle back to several of these points with more detail if people are interested.

CO-CHAIR DAMBERG: Thank you so much for that great summary. And also appreciate some of your additional analytic work to address some of the questions that were raised in the pre-review.

What I'd like to do now, is turn to our reviewers. And Jennifer Eames Huff is first up to review importance. Jennifer?

MEMBER EAMES HUFF: Hi. I'm going to say this morning Larry did a great job of setting

the context of the importance of cost and resource use. So, I'm going to jump directly to the measure.

So, in terms of high priority, to demonstrate this, the developers' site data, I think which we are all familiar with, the percent Medicare expenditures account for the gross domestic product as well as how much hospital costs account for Medicare expenditures in total.

The staff review scored this as high. And the comments were overwhelmingly in agreement with the review.

In terms of performance gap, they used 2015 data. And the Medicare spending per beneficiary, the hospital measure score was .99 with a standard deviation of plus or minus .99.

The provider score ranges from a minimum of .59 to a maximum of 2.25. The median value was .99. And they had the 25th and 75th percentile of the measure score as .94 and 1.03.

They also did a similar analysis for the cost piece. So, the mean -- Medicare

standardized cost. Mean Medicare spending per beneficiary, the hospital amount was approximately 20,000 with a standard deviation of 1800 dollars.

The provider amounts range from a minimum of approximately 12,000 to a max of 46,000. The median was 20,221. The 25th and 75th percentile amounts were around 19,000 and around 21,000.

They also did an analysis of the provider score changes between 2014 and 2015. And the results showed that 47 percent of hospitals improved on their measure score. Which is defined as having a lower score.

In terms of disparities, I think we just -- it was focused on the SDS risk adjustment. And I think we just heard the measure developer summarize that. So, I'll skip over that and say the staff review was scored as high.

In general the comments were supportive of this particular area, I think.

There's overall agreement, but there were some caveats. Concerns were raised with regards to the opportunity for improvement and the challenge for hospitals to affect change in what this measure covers.

And there were also some critiques about the SDS. Which I also think we heard some of that in the public comment.

The intent of the measure is to incent hospitals to coordinate care and reduce unnecessarily -- unnecessary utilization during the episode. This is also scored as high. And I didn't see comments on this in the updated measure sheet.

CO-CHAIR DAMBERG: All right. Thanks Jennifer. Why don't we start by seeing if anyone has any questions, comments in the room before we turn to the phone.

(No response)

CO-CHAIR DAMBERG: All right. Seeing none in the room, any folks on the phone have questions?

(No response)

CO-CHAIR DAMBERG: All right. Hearing none, I think we can move to a vote.

MR. SINGH: Okay. So we're going to begin voting on Measure Number 2158 regarding a high priority. Your options are one high, two moderate, three low, or four insufficient.

Please begin voting.

(Voting)

MR. SINGH: Okay. So all votes are in and voting is now closed for high priority for Measure Number 2158.

The results is as follows: 18 votes for high, zero votes for moderate, zero votes for low, and zero votes for insufficient. That accounts for 100 percent of votes for high.

And this measure passes for high priority.

CO-CHAIR DAMBERG: All right. Thank you. Right, we're now going to shift to what I think is going to be a little more challenging space. Dolores Yanagihara has offered to take us

through the scientific acceptability.

MEMBER YANAGIHARA: Offered?

(Laughter)

CO-CHAIR DAMBERG: You're a really good volunteer. I'm sorry.

CO-CHAIR ASPLIN: Cheryl, could we just take the remaining --

CO-CHAIR DAMBERG: Oh, I'm sorry.

CO-CHAIR ASPLIN: The remaining votes for gap and intent. And then we can move onto science.

CO-CHAIR DAMBERG: Okay.

CO-CHAIR ASPLIN: So we've got two more just sub-criteria votes here.

CO-CHAIR DAMBERG: Sorry. I'm kind of new at this game.

CO-CHAIR ASPLIN: No, no. It's okay. I think everybody wants to get to that section for discussion.

CO-CHAIR DAMBERG: All right. So, the next section is, I'm going to need your help.

CO-CHAIR ASPLIN: It's 1B, gap in

care. So, Irvin, you want to just walk us through the voting slides, please?

MR. SINGH: Sure thing. So, we're going to be voting on gap in care/opportunity for improvement for Measure Number 2158. Your options as follows: one high, two moderate, three low, four insufficient.

Voting has now begun.

(Voting)

MR. SINGH: Looks like our denominator is off. So, we might need to start voting one more time.

All right. So we're going to start voting again for gap in care/opportunity improvement for Measure Number 2158. Please submit your votes.

(Voting)

MS. O'ROURKE: So we have the phone votes, right? So folks on the phone, you are okay. Folks in the room, we're going to clear the slide and vote again.

We're having the same issue where it

seems to have only captured the high votes. So, if you could vote one more time.

MR. SINGH: Vote now.

MS. O'ROURKE: I'm sorry, vote now.

(Voting)

MR. SINGH: Okay. So all votes are in and voting is now closed for gap in care/opportunity for improvement for Measure Number 2158.

The results is as follows: 12 votes for high, six votes for moderate, zero votes for low, and zero votes for insufficient. That accounts for 67 percent of the votes for high, 33 percent of the votes for moderate, zero percent of votes for low, and zero percent of votes for insufficient.

This measure passes for gap in care and opportunity for improvement.

CO-CHAIR DAMBERG: Great. So we have one more component. And that is measure intent that we're going to vote on next.

MR. SINGH: Okay. So we're going to

be voting on measure intent for Measure Number 2158. Your options are one high, two moderate, three low, and four insufficient.

Please submit your vote now.

(Voting)

MR. SINGH: So it looks like our denominator is off. We had a couple of phantom votes. So, we have 20 this time. I don't know how that happened.

Just bear with me while I restart the vote. Please submit your vote now.

(Voting)

MR. SINGH: All right. So all votes are in and voting is now closed for measure intent for Measure Number 2158.

The result is as follows: 13 votes for high, five votes for moderate, zero votes for low, and zero votes for insufficient. That accounts for 72 percent of the votes for high, 28 percent of the votes for moderate, zero percent of the votes for low, and zero percent of the votes for insufficient.

This measure passes for measure intent.

CO-CHAIR DAMBERG: Great. Okay, so let's go back to scientific acceptability. And Dolores, do you want to walk us through the components? Thank you.

MEMBER YANAGIHARA: Okay. So, we'll start with reliability. So the numerator for this measure is basically risk adjusted average spending. So, observe to expect tons of national average spending.

The denominator is episode weighted median spending nationally. And the exclusions are zero dollar claims, transfer -- acute to acute transfers, those hospitals not paid through IPPS, the top and bottom one percent, those patients who may have incomplete data, so part A only, became deceased, Medicare Advantage, or Medicaid Primary.

And then certain hospitals with -- that focus on psychiatric rehab, pediatrics, or primary cancer or research focused.

Some of the -- one of the issues raised with kind of the specifications was the disability code used. There was a suggestion that using the OREC code, the original reasons for eligibility code, would have been preferable.

And then there was one comment about kind of a concern about just basing performance on DRGs is limiting. It's not kind of the full scope.

In terms of testing for reliability, there was new data that was provided. The data was from all of the hospitals for 2015 from January through December 1.

There was no real testing done on the data element aside from the regular CMS testing that's done, the auditing that's done and data analysis that's done to look for data completeness. But at the measure score level there was testing done.

There were two different kinds of test. Well, two different kinds of tests and then two variations within one of the tests.

So there was a test/retest done. The first one was done between 2014 and 2015. And for that test/retest, about three quarters of the hospitals that were high or low, were steady between those two periods.

And the correlations were -- both a Spearman correlation and the Pearson correlation were both very high. Over a .8.

There was also a test/retest done with random subsets within 2015. That one the agreement between the subsets was a little bit lower, but still over 70 percent of the high or low quintile hospitals stayed in the same quintile.

And again, the Spearman rank correlation was high, .82. Pearson was a little bit lower, .7. But that was expected.

The other test that was done was a reliability test. And that was done for hospitals with greater than or equal to 25 episodes.

For -- and I think that Sriniketh

brought some additional information that I'll try to incorporate in here. But, 68 percent were over a .9 reliability.

And then he added that 93 percent were over a .7 reliability. Which means there would be 7 percent that were, you know, less than .7 reliability.

Ninety-nine percent were above .4. Which was the threshold that you had originally used. Some -- the concern -- the basic concern raised was the .4 reliability and you adjust that.

So, I think that was my notes, what I had for reliability.

CO-CHAIR DAMBERG: All right. I'm going to open it up for questions, comments. For those on the phone, please raise your hand and I'm going to start with Nancy in the room.

MEMBER GARRETT: So I have a question about risk adjustment. Is this the right place? Or is that validity?

CO-CHAIR ASPLIN: That's validity.

MEMBER GARRETT: Okay.

CO-CHAIR DAMBERG: Bill?

MEMBER WEINTRAUB: Hi, so I have a question for the developer and for CMS. So, at a -- really at all the hospitals this looks reliable.

But what can we say about the individual hospital? Because we're looking at thousands of hospitals. And some hospitals, and this is more -- part of my comment this morning, similar.

Some hospitals may stand out on a stochastic basis either looking poor or looking very good when it's inappropriate. Or just looking like everybody else when really they really should be -- they really are an outlier, just on the basis of the number.

Especially for some of the hospitals where the numbers are small. There are various approaches to that. None of them are entirely satisfactory.

And I just wonder the thoughts of the

developer and CMS?

DR. NAGAVARAPU: So real quickly. On the original disability point. It wasn't noted in this mission form. But we did use the original reason for disability.

So, that -- yes. And so the disability flag, wherever that's expressed, is the original reason for disability.

CO-CHAIR DAMBERG: Actually, can I follow up on that? Because that was my comment.

DR. NAGAVARAPU: Yes. Sure.

CO-CHAIR DAMBERG: So did you guys use the enrollment database or the integrated data repository?

DR. NAGAVARAPU: It's from the enrollment database.

CO-CHAIR DAMBERG: Okay. So the reason I raise this, because we've been, in work that I do for CMS, been using that same variable. And it turns out that at age 65 it resets to aged into Medicare.

Whereas the variable that's in the

integrated repository, it's the OREC variable. It retains their disability status beyond age 65. So, it's an easy tweak for you guys.

But, I just wanted you to be aware of that. So if you did your cost tabs, you wouldn't find anybody over the age of 65 who is disabled.

DR. NAGAVARAPU: And just to make sure about that, there is a current reason for eligibility that's also included in EDB. And we're not using that.

So, we're not using current disability. We're using EDB disability.

CO-CHAIR DAMBERG: Yes. No, you want the OREC. But you want it from the integrated data repository.

And I don't know sort of all the ways Medicare data flows. But somehow or other it doesn't get reset in the IDR. It's kind of strange.

DR. NAGAVARAPU: So, on the other comment, so we did some additional numbers just to get a sense of how much hospitals are moving

around in the type of comment that's being made there.

Because the reliability numbers reflect that in part in the sense that we're calculating reliability. Kind of a signal to the within metric for each hospital.

And trying to understand how much the within variance is relative to the between variance, and I think that's a point that maybe Dr. Needleman and Dr. Naessens had brought up in the last session. And so the reliability metric speaks to that a bit directly in terms of how much hospitals can be distinguished from one another.

We've also done some examination of quintile rank stability that was mentioned in the summary by Dolores. We looked into that a little bit more. We would expect some movement across quintiles across years. And actually that's desirable.

We want hospitals to be able to improve on the measure. And so we would expect

rankings to change over time.

But, we also think there should be some stability and not just noise. That we're distinguishing something about these hospitals. And we looked into one number where 94 percent of hospitals that are in the highest quintile of spending in 2014 are in the highest quintile files in 2015.

So while there is some movement across quintiles, the limited amount that Dolores mentioned, there's not drastic movements across time.

And then we have some information on some simple calculations of confidence intervals. I know some people asked about this.

We did some simple calculations. And for example you can distinguish sort of high performers from other hospitals.

So, about 75 percent of providers can be statistically distinguished with a 95 percent simple calculation of a confidence interval from the 10th percentile. So, from high performing

providers. To give you a sense of that.

CO-CHAIR DAMBERG: Okay. So -- sorry.

MS. SPALDING BUSH: Can I -- oh, sorry.

CO-CHAIR DAMBERG: That's okay.

MS. SPALDING BUSH: This is Kim. I hope you guys can hear me. I would also just add that we did set the case minimums in the interest of protecting those smaller hospitals from having one or two cases that, you know, sent their score too far in one direction or another to protect the stability there.

I mean, we also do a -- and Sri can speak better to the actual way that we do this. But we address outliers.

So we take a look at the difference between the expected cost and the observed costs. And when they're way off, they do dial that back a little bit to help mitigate some of the impact of those outlier payments for certain pieces.

But I think would also potentially

have a bigger impact on a smaller hospital.

CO-CHAIR DAMBERG: All right. Thank you. Jack?

MEMBER NEEDLEMAN: This comment/ question sort of straddles the reliability/ validity dimensions. So, I apologize if I slide into validity.

But, I think it's important to -- as we think about the reliability here, to understand what's being measured. And therefore how we ought to be testing the reliability or the validity.

This is fundamentally a price-weighted measure of utilization. And because DRG is included in the risk adjuster, all the differences in prices that are associated with the DRG are sopped up in the risk adjuster.

The DRG may also pick up any differences across DRGs in the amount of non-hospital use that is associated with specific DRGs. But it ought to be sopping up the price, the standardized price that is in -- that is part

of this estimate.

So, given that it's doing that, what the measure is fundamentally measuring is all the variations in non-hospital care associated with that hospitalization.

And my -- so, given that, and given the role of the DRG and the risk adjuster, I'm wondering, have you done any analysis of the reliability or any of the other measures of variability taking the DRG price out of the measure?

So we're really looking at just the variability and the non-hospital components of use. Which is really what the measure is capturing in terms of the variability.

DR. NAGAVARAPU: Thanks for that. So, that's exactly right. That DRGs used in risk adjustment sort of net out the standardized cost of DRG.

There is still some variation from the inpatient stay in terms of Part B costs. So, the extent to which physician services are used and

so on, will have an impact.

In terms of the results there, the fraction of variation accounted for by that is on the order of 15 percent. Whereas the fraction accounted for by post-discharge spending is higher.

What we have done is some basic analysis as to where the post-discharge spending is coming from. And to get a sense of what's sort of driving the variation in that to speak to the point about, you know, just focusing on post-discharge. And it seems like a lot of that variation is coming from particular post-acute care facilities. So, it's skilled nursing facility intensity of use, for instance, and home health.

But we totally agree that the post-discharge portion of this is a key aspect of understanding what's driving the variation in performance across providers. And there's certain key post-acute care facilities rather than later readmissions, for instance, that seems

to really be driving a lot of that variation.

CO-CHAIR DAMBERG: And I'm going to turn to folks on the phone. John, do you want to start?

MEMBER RATLIFF: Yes. Just a quick question. You note that you use 25 episodes for your reliability testing. That's kind of your cutoff, like each facility having to have 25. And that you also looked by restricting the number of facilities.

So this is where you have 50 episodes. And you note that you drop the number of hospitals. And then you don't really give us like a real -- reliability, excuse me, score, when you increase the number of episodes.

And I was wondering what you saw with that? Or if the developer would just like to comment on that?

Like how much of an increase in your reliability you reached by kind of decreasing the number of hospitals? By going to hospitals with a higher number of episodes?

DR. NAGAVARAPU: Yes. Thanks for that. So, I have the numbers right here in front of me actually that we can send everyone afterwards.

But, if you look at the .7 threshold, if you use a threshold of five episodes, reliability is at 94 percent. And are between 93 and 94 percent for a lower episodes.

And then it stays fairly stable. So, you know, it stays at 93 percent all the way through to a case minimum of 75 episodes.

And it doesn't really show an increase until you get to about 110 episodes as a case minimum. Where it increases only slightly to 95 percent.

That's for .7 reliability. You know, for .4 as well as .8 and .6, it's also very stable across case minimums.

And I think what that's telling us is that, you know, the measure is picking up something real that's not sort of swayed very dramatically by the exact sample that's being

included here.

CO-CHAIR DAMBERG: Bill, I think you're next in the queue.

MEMBER WEINTRAUB: Did you say Bill? I'm not listed as one of the people to be speaking right now.

CO-CHAIR DAMBERG: Okay. Thank you. Maybe your hand was still up on the site.

MEMBER WEINTRAUB: Oh, I never lowered it. Sorry.

CO-CHAIR DAMBERG: Okay. Thanks. I have one question regarding the reliability. So if approximately 7 percent or below .7, does CMS intend to use the measure results if it falls below that reliability threshold?

MS. SPALDING BUSH: Yes. I think for all of our payment and quality measures, we do set the case minimum. And then we, you know, kind of place where generally hospitals are meeting an acceptable standard of reliability.

And they do use the measure then for all hospitals that meet that case minimum.

CO-CHAIR DAMBERG: Any other comments on this before we turn to talking about validity?

MS. SPALDING BUSH: Can I just add to that? I'm sorry.

CO-CHAIR DAMBERG: Sure.

MS. SPALDING BUSH: It's hard to hear. I'm really sorry. We also, you know, we split that balance between being inclusive, capturing enough hospitals. And then setting a threshold so high that hospitals start to fall out.

So, I think that's where we -- I think that's where we try to land. In that place where we've got a reliable measure. Where we've not had a whole lot of hospitals drop out. Especially in something as important as measuring costs.

CO-CHAIR DAMBERG: Great. Thank you. Any other comments before we move to validity?

(No response)

CO-CHAIR DAMBERG: Sure. All right, before everybody forgets what we just discussed, let's take a vote on reliability.

MR. SINGH: Okay. So we're going to begin voting on reliability for Measure Number 2158. Your options are one high, two moderate, three low, and four insufficient.

Voting has now begun.

(Voting)

MS. O'ROURKE: Larry, if you could submit your vote in the comment box, that would be great. Thank you.

MEMBER BECKER: Did you get it that time?

MS. O'ROURKE: Nope. Still not seeing it. If you want, you can tell me right now if you're comfortable doing that or you can email it to me.

MEMBER BECKER: Reason it's not working. Well, one.

MS. O'ROURKE: Okay. Thank you.

MR. SINGH: Okay. So we're going to try that one more time because the other votes did not come in.

So, bear with me while I reset it.

And I'll let you know when to resubmit the vote.

MS. O'ROURKE: Folks on the phone, we have your votes. So, it will just be people in the room will need to revote.

Apologies. We're having the same issue where it's only capturing highs.

MR. SINGH: Okay. So, please resubmit your vote again for those that are in the room.

(Voting)

MR. SINGH: Okay. So all votes are in. And voting is now closed for reliability for Measure Number 2158.

The results is as follows: eight votes for high, ten votes for moderate, zero votes for low, and zero votes for insufficient. That accounts for 44 percent of the votes for high, 56 percent of the votes for moderate, zero percent of votes for low, and zero percent of votes of insufficient.

This measure passes for reliability.

CO-CHAIR DAMBERG: All right. Now we can turn to validity. Dolores?

MEMBER YANAGIHARA: All right. So, the first part of validity is the specifications. And I think there was a general sense that the measure specifications were consistent with the intent of the measure.

There was one question raise, or concern raised about the ability of hospitals to influence the total amount. Especially, I think, the post-acute care piece.

In terms of testing, there was new data that was used, empirical testing at the measure score level.

The first test was looking at the correlation with other measures of spending and service utilization. And the second testing was looking at cost variation by time period within the episode.

So in terms of correlations, the first one was correlated with risk-adjusted per capita spending at the HRR level. The correlation there was greater than .5. With a range, I believe of like .51, .61 or something like that. I think

there was a range there.

And then the correlation with post-acute -- not SNF, well I guess skilled nursing and inpatient service per episode was .52. So, moderate correlations there.

Variation by time period, 84 percent of the variation was accounted for by post-acute care. And 11 percent by the three days prior and the in-depth admission length of stay.

So, there was a concern raised about why you use different tests this times then the first time around. Which I think focused on the condition specific measures of utilization.

And there was a question raised about whether a moderate correlation with the other measures of utilization and spending were sufficient. If they should be stronger.

And then again, the question of being able to control the post-acute care. Because that was where most of the variation in spending occurred.

And then there was one question

raised, it may be more of a spec question. But anyway, it was about in certain situations, I think that's IRF, which I'm not sure if that's inpatient rehab facility, I was guessing. Okay.

So, that that sometimes that can actually extend beyond 30 days if they're in the IRF longer then the 30 days after the whole amount is included. And so there was concern about that and the impact of that.

In terms of threat to validity, there was some testing done on exclusions. It was found that the -- there was minimal impact and high correlation coefficients. Let's see, there was -- I think, okay, so it's 1.6 percent of cases excluded due to acute transfer. So that's not too much.

And a high correlation, .95. About eight percent of episodes were excluded due to death during the episode. Very high correlation with and without those of .99

For overlapping episodes, that accounted for about 12 percent of exclusions.

And again, a very high correlation of whether you include them or exclude them of .99.

Removing the top and bottom outliers, there was a -- that was the one -- top and bottom one percent. There was a correlation of .93. So again, quite high correlations whether you include or exclude those populations.

Some of the concerns raised were -- someone raised a concern about cancer hospitals being excluded. But there may be some general hospitals that have a lot of cancer patients that might be included and that, you know, how would that impact them? And that was the main concern on exclusions.

In terms of risk adjustment, the risk adjustment is based on the CMS-HCC model. But there's no adjustment for sex in this one.

Let's see, it includes age disability, ESRD status, long-term care status, severity of illness, and the MS-DRG of the index submission. And then ordinary least squares linear regression is used.

In terms of discrimination, the R squared for -- across all MBCs was .3. And then overall was .48.

In terms of calibration, they looked at -- there was a comment about the 90-day look back period, and was that sufficient? So they looked back 365 days to see what affect that had.

And basically the R squared of .3014 went to .2997. So, basically no impact there. But it did cause 6.7 percent of the episodes to be excluded because of the longer look back period.

We looked at adding institutional status. And there was minimal improvement to the R squared. But it decreased the number of other variables that were statistically significant. And so that was not included.

Terms of predicted ratios, I think Sri talked about this. By risk decile is very consistent in all the deciles.

And then in terms of SDS testing, they used income to poverty, and also race: black/non-

black. The F test showed both significant predictors, both of those were significant predictors.

But when they added it to the risk model, it didn't have much impact at all. So, it was a plus or minus .01 for 97 percent of the hospitals. And nearly 100 percent correlation.

There were a lot of concerns raised about the SDS. I think we heard some of that from AAMC.

And then there was a question about using zip code level. I think Sri explained that the reason for doing that was to kind of find a balance between individual income impact and community income impact.

And let's see, oh, there was a question about who were the three percent that were affected more than .01 percent or .01. And whether those were certain categories. For example, like rural hospitals or something like that.

And so I don't know if you have more

information about that. But if you do, that would be great.

And then people raised the issue with the connection to quality. And so the point was that sometimes there's more services that are provided that can actually lead to better outcomes. But we don't know what the outcomes are. In this case, it's just spending. So, there was a question about that.

I'm almost done. This is a long section.

(Laughter)

MEMBER YANAGIHARA: In terms of meaningful differences, they stratified results by hospital characteristics. So geography, teaching hospitals and rural or urban. And the results were as expected based on literature.

There was a request by a member, or more then one member maybe, that they would like to get information on the source of differences for the higher and lower cost hospitals.

So, kind of understanding that more.

Like what was really driving that? And then, again the link to quality was brought up.

In terms of missing data, I'm not sure if this is actually missing data or not, but two comments were that there's no Part D data. So, that is true.

And then Medicare Advantage is not included. And so you know, there may not be missing data for Medicare fee for service, but in terms of like understanding the context and the whole market, that Medicare Advantage would be missing.

So, those are my notes.

CO-CHAIR DAMBERG: Terrific. Yes, Taroon?

DR. AMIN: Before we get into conversation on this, Dolores, I think you did a great job by the way. There were a few pre-meeting comments that were submitted related to this measure. As particularly related to this section, that I just wanted to highlight.

We heard from the AAMC on this, from

the federation. There were some questions around the variable that was used in the SDS evaluation as it relates to the ASPE report.

And then from the AMA, some questions around the level of analysis. And concern about using this measure at a physician level. But it's not specified as such.

So those are the only other two comments that are related to validity that came up. I just want to throw that into the discussion.

CO-CHAIR DAMBERG: All right. So why don't we start with Nancy.

MEMBER GARRETT: So I have a couple of initial comments here. One is, to me the major validity concern here is actually the title of the measure.

I brought this up the last time we reviewed it. And I'm sure when something's written into statute, we can't change the title.

But, to me it's just ruining this thing. I mean, this measure should be called

something like resource use per hospital episode. It's not a true cost measure. And it's certainly not a per beneficiary measure.

It's per episode. And so it just seems really misleading. And given that, you know, this is available to the public and it's meant to be used by consumers, it seems really important that the title is accurate.

So, I think that's a real validity concern that the title doesn't match what the measure is.

And then the second thing I wanted to bring up is something that we noticed using the data in Minnesota. Which is that there's a real bifurcation between rural and urban hospitals on these results.

And we find that the hospitals in the more urban settings in the state have higher scores. And then rural have lower scores on this measure.

So like in Minnesota for example, the Minnesota average is .9. And every single urban

hospital is .94, .95, .96. And every single rural hospital is under .88.

And so I actually corresponded with the measure developers about this a couple of years ago. And one of their responses was, well there's research showing that urban hospitals use more resources.

And so, that could well be the explanation. But it also makes me worry that maybe there's something about the risk adjustment that isn't picking up tertiary care and some of the complexity of patients referred into some of these facilities in the urban areas.

So, I don't -- I haven't done that analysis across other states. So I don't know if that is something that you've seen other places. But to me, that's a real concern about how this measure is actually playing out.

CO-CHAIR DAMBERG: Did you want to comment on that?

DR. NAGAVARAPU: Yes. So, I'll let Kim respond to the title issue. But the

rural/urban. On average rural hospitals do have lower scores than urban hospitals.

I think in the materials that you all saw, it sort of scores nationally. And the mean for urban hospitals is 1.0. Whereas for rural hospitals is .95.

However, the one thing I want to point out, is if you look at the percentiles and the distributions of rural hospitals and urban hospitals, the distributions overlap a fair amount. Right?

So, I wouldn't think of this as a measure that's somehow segmenting off for urban and rural hospitals. There are a large swath of urban hospitals that are performing better then the large set of rural hospitals.

So, I think what you're seeing in terms of the difference in average is true. Although the size of the difference seems to be larger in Minnesota than nationally.

But those distributions are overlapping a fair amount. So, I do think

there's something exclusive about the evidences that's preventing urban hospitals from having this source.

MS. SPALDING BUSH: So it -- oh, I'm sorry.

CO-CHAIR DAMBERG: No problem.

MS. SPALDING BUSH: Just regarding the title of the measure, you're right. It was written in with the statute and for inclusion in the Hospital Value-Based Purchasing Program.

When it displayed on Hospital Compare, it does have a different name. Because that site is for beneficiary consumption, theoretically although a payment measure maybe of more interest to others as well.

But, it's called like Medicare -- or spending per hospital patient with Medicare, I think. Something like that. It gives you the sense that it is a hospital related measure and it has to do with payment, so.

CO-CHAIR DAMBERG: Brent?

MS. SPALDING BUSH: We had the same

feeling about the title in CMS that you did.

CO-CHAIR DAMBERG: Brent?

CO-CHAIR ASPLIN: My question is about the risk adjustment model and the 90-day look back period. Could you just comment on that?

Is it -- either there's a benefit of a look back period, in which case I don't think 90 days is enough if you apply how the HCC model works for risk adjustment in Medicare Advantage. You would never just look at a quarter. Or you would miss a ton of risk in your population.

Or there's really no benefit in the look back period and you're going to capture whatever you're going to capture in the three days prior to the episode. And you'll get your HCC diagnosis there.

Would you just speak to that? Because the 90 days to me just seems hung in the middle. Sri?

DR. NAGAVARAPU: Yes. No, that's a good question. So when we developed this, we tested a variety of look back periods. As you

noted, Medicare Advantage looks at the previous year.

What we did is look at sort of 90 days, 180 days, and so on. And up to 365 days, to get a sense of how predictive power changes.

And the trade off there is potentially improvements in predictive power versus losing more beneficiaries because of people's move, they get out of Medicare Advantage, you may lose people the further you look back based on the restrictions we have. Right?

However, what we found in the analysis here, and what we found in previous analyses as well, is that moving to a further look back actually slightly reduces the R squared in moving from 90 days to 365 days. And we think what's happening is a problem with sort of false positives in the sense that there are some conditions that you would pick up if you look 365 days back that resolve by the time you get to the inpatient stay.

And we think that's one possible

explanation for what we're seeing. That the 90-day look back actually performs a bit better in terms of predictive outcome.

But we have looked at a range of -- when developing the measure for exactly the concern you have.

MS. SPALDING BUSH: Yes. And the difference there being it's predicting just the cost of this episode.

So that per -- it maybe that those 90 days pick up the conditions that are most directly impacted, you know, the reason that person was admitted, their, you know, their current state at the time of admission.

And what we would expect the cost to be for them post-discharge as a cur -- you know, as compared to things we just pick up that they happen to have as an acute condition that hasn't been treated lately, you know, if we went back a year.

CO-CHAIR DAMBERG: All right. I'm going to turn to Jack. And I think we need to

try to wrap this up soon. So, --

MEMBER NEEDLEMAN: I actually have just a few comments. I think the measure is, as much as we looked at it originally, and it's as worthy of endorsement as the first one was.

I think there are a couple of things that should be noted here. Not necessarily for the development of the measure. But for your thinking about what you've got and how to analyze it.

You show a fair amount of variability, like 25 percentage points in your index between the -- you know, the 10th and 90th percentile. And that's a lot.

But it's even worse than that, because you have standardized out about half the costs, which are the hospitalization costs. So, the variability around the non-hospital costs are much larger.

And I think it's very important for usability purposes, as well as for interpret ability purpose that you do a lot more analysis

on where that -- what those variations in spending are.

You mentioned readmissions as one of the major potential sources of that variability, and SNF use as the second major source. And we need to understand that a lot better. We need to understand that particularly if it's going to be used in a value-based payment system, so we're also going to be correlating with the outcome.

So, I would encourage you to think hard about that. The second thing again, just a general comment.

This is not a measure of resource use because of the standardization of prices. To the extent that you've got individual visits or individual days in the SNF there, yes, you've got some measure of differential resource use.

But within a hospital, you don't have any variation in resource use. Because you've used a standardized price for it. A little bit on the Part B, but only a little bit on the Part B.

And if we're going to understand how hospitals influence those other costs, we better have a better understanding of what's happening in the hospital then this measure provides.

With regard to risk adjustment and SDS risk adjustment, since Jan isn't here, let me channel her. When your Yale colleagues were here presenting some other measures, we raked them over the coals over using zip code level measures of income rather than census track level measures.

And we were told in subsequent work, it's just really hard to get the beneficiaries matched up to their census tracks. Even though we supposedly have the address stuff.

So, it was not feasible. We just had a measure, two measures earlier today in which health partners managed to match their patients up with census tracks.

I don't find the argument that the zip code better captures the full range of resources available, either to the patient or in the

community, as compelling as understanding better the local circumstances in which the beneficiary lives. Particularly if you're trying to apply that as a measure of what resources the beneficiary has.

So, I want you to -- I want to strongly encourage you to go back and reconsider the SDS analysis. We talked about the other problems of, you know, high disparity, you know, high disadvantage using facilities which are not captured in your risk adjuster either.

But, I really do want you to go back, if we're going to be doing risk adjustment around SDS, you've got to figure out how to get the -- and you're going to use some geographic proxy for the circumstances of individual beneficiaries.

You really do have to go back to the census track as the basis for that calculation, not zip codes.

CO-CHAIR DAMBERG: Nancy, did you have one last question before we wrap up?

MEMBER GARRETT: Yes. I just wanted

to second that comment. I feel like the approach here to risk adjustment for SDS is really not adequate for where we need to be.

And especially given the really big impact that readmissions has on this particular measure and all of the discussion that's been going on around readmissions and the strong link to socioeconomic circumstances of the beneficiary.

So, I think that there's more we need to do here.

CO-CHAIR DAMBERG: And jut to follow onto that. I guess I was a little confused. And maybe you could say more about the analysis you did, adding dual status.

And it sounds like your results differ a significant amount from what the IMPACT Act analytic group found. And wondering if you could say why you think that's the case?

DR. NAGAVARAPU: Thanks. So, let me start with that point. And then I'll circle back to all the other points.

So, on dual status. So we're very familiar with the ASPE report. I came in, actually was the support contract period to do the analysis for ASPE. So we know those results very well.

If you remember from the report, the key distinction that they're making in the report is a comparison between safety net hospitals and non-safety net hospitals. Where safety net is defined as greater than 20 percent dish.

When they look at the magnitude of difference in measure scores in 2015 and fiscal year 2015 between those two groups, the difference is .01. And so the difference is relatively small. Where we think of the standard deviation here as .09.

So, that's the first point. I think there is a question as to what is driving that .01 difference between safety net hospitals and non-safety net.

And so we did some analyses to look into it that the ASPE report didn't get into.

And basically what we did is, we divided up hospitals by the fraction of their episodes that are in the -- that are for dual beneficiaries.

And looked at different thresholds to compare hospitals that we would classify as heavily dual. Like above a certain threshold. Versus non-heavily dual.

Where the heavily dual are going to correspond to some extent to the safety net hospital. Because we want to understand what's driving that .1 in the ASPE report.

And what we -- .01, sorry. And what we saw in that, in the discrepancy there, is we looked at the -- we calculated the measure for non-duals in those hospitals.

And what you see is that even on non-duals, the hospitals perform worse on the MSPB measure then in the high dual concentration hospital, as compared to the low dual concentration. And so if you look at the number, I have it written here. That's right.

So if you use the threshold of greater

than 50 percent for dual episodes as the marker for what's a dual hospital and what's a non-dual hospital, and then calculate the MSPB measure for just non-duals in both of those fee sets of hospitals, that measure is .04 larger for non-duals in hospitals with greater than 50 percent dual episodes versus others.

So, what that's saying is that both for duals and non-duals, these hospitals seem to be more expensive. It's not something that's driven specifically by dual status.

Now, it could be some sort of spillover type affect that everyone had mentioned. That's a possibility. Another possibility is that these are just inefficient providers.

And so I think it's something that we really do need to dig down into. I think more research on that is required.

But, that seems to be what's driving this result in the ASPE study to some extent. And that's something we could dig into more with

the safety net definition.

For -- in terms of clarifying what we did with dual status, we understand the concern. The concerns about census track. We thought zip code is a good way to get at sort of both of these concerns at the same time. About the individual level and the community level attributes.

But we recognize that the people wanted to focus more on the individual level. That's why we went straight to dual status to look at this.

And what we did is use a dual dummy in the risk adjustment model for the MSPB measure. Calculated the MSPB measure both with and without the dual dummy and look at changes in the measure.

And saw that greater than 98 percent of hospitals have a change in magnitude less than .201, so.

CO-CHAIR DAMBERG: So, I think the challenge with risk adjustment is, you know, for

most providers, it's probably not going to make much difference. But it's always, there's sort of a handful of providers where the difference can be quite large.

And I guess the question is, is have you looked at that in terms of how much impact it has on those that have a really large fraction say of duals?

DR. NAGAVARAPU: Yes. No, that -- sorry, thanks. One way we've looked at that is first to just see how many hospitals are sort of moved by -- to sort of large changes in MSPB measure score by including SDS factors or dual in risk adjustment.

Looking at the table of numbers for SDS, you know, we said the number that it's something like 98 percent or 97 percent or so, within .01. But, if you look at the other hospitals, it's very rare that any hospital is above .1 in the change in either direction.

There's only one hospital that's above .03 and one hospital that's below minus .03 in a

change. So it's very tightly confined. For duals it's similarly tightly confined in terms of the numbers. Even the min and max are quite tight. And I can pull the data up for you in a second.

So if you look at the .1 percentile, the change is minus .016. And if you look at the 99.9 percentile, the change is .029. So it's not huge fluctuations for hospitals. It's like minor fluctuations.

CO-CHAIR DAMBERG: All right. Thank you. Sure, Helen?

DR. BURSTIN: I think there's been so much attention to the ASPE report, I just want to clarify one additional thing.

So, the actual table from the ASPE report points out a four percent difference. Is that what you're talking about, the .04? Four percent difference using dual beneficiaries.

And then they specifically then looked at this question of medical complexity using a claims-based frailty index. Planning that half

of it's potentially addressed by in fact duals isn't just poverty of course, it's for functional status, it's disability, et cetera.

And so they then found about half of that was taken up by these potentially unmeasured clinical complexity. So at least their recommendation was that this should potentially be adjusted for.

So again, even if it's a small effect, I just want to clarify the perspectives of both CMS and the developer. Just because again, there's this report that's now fresh out on the street saying something that's slightly different.

DR. NAGAVARAPU: Yes. So, the table I had in mind for the .01 for the fiscal year 2015 is further down in the report. And I think it's in the section where they talk about policy simulations.

For this table they showed the four percent total effect for social risk. What this table is not doing that sort of the supplementary

results that we have help confirm, is that that four percent could be something that's driven by the types of hospitals that dual eligible beneficiaries are in, versus themselves.

And so the four percent sort of includes both of those effects. Now, they try to get it that with that second column there with the within hospital.

But if you look at what they do for within hospitals, it's a random affects model. And so it doesn't actually net out the mean within a given hospital.

All it's doing is making adjustment that's basically relating to sort of the clustering. Taking into the account the air structure.

And so really what they would want to do in that second column to try and see how much is due to which facilities people are concentrated in versus the dual individuals themselves, would be around like a fixed event sort of specification. Which they don't have

here.

Now, the analysis we did that sort of tries to look at these sorts of hospitals where duals are more concentrated and look at what's happening for non-duals, does provide some indicative evidence that there's something going on here. Where the types of facilities that duals maybe clustered into, are facilities that are inefficient not only for duals, but also for non-duals in this measure.

DR. BURSTIN: Right. But part of what this analysis is suggesting is that dual status is an indicator of clinical complexity. So inefficiency could be related to under-measure clinical complexity, I think, is what the report is urging. As you can see.

DR. NAGAVARAPU: Yes. No, and I think it's possible. I think the HCC model that we have now does a good job of taking into account a lot of clinical factors that may affect dual status.

And that's why I think that the MSPB

measure doesn't do that much when we risk adjust for dual. Because a lot of those factors are already captured by the existing HCC flags.

To the extent that there is additional clinical complexity that's involved there, that's a possibility. I think what goes hand in hand with this decision though, is a policy decision about, given these sort of very small size effects, do we actually want to create a separate standard for dual eligible beneficiaries?

Or do we create a lower standard of care essentially by allowing that to enter into this? I'm not sure.

CO-CHAIR DAMBERG: So, can I actually comment on that?

DR. NAGAVARAPU: Yes.

CO-CHAIR DAMBERG: So, we also, full disclosure, did some analysis, not on this specific measure. But, one of the things that we were asked to look at for the IMPACT Act was within provider variation on a particular measure.

And I'm curious for, let's say dual versus non-dual within these hospitals if you compute the same measures. Like how much disparity is there? And is it sort of large and consistently negative?

So, I think that's type of analysis, if you haven't done it, and I'm happy to share with you work that we did, that might be informative in terms of your decision making around adjustment.

DR. NAGAVARAPU: Thanks for that. Yes, so we actually do have some results comparing them.

And if you use that 60 percent threshold in the preliminary results that we had in order to respond to some of these pre-evaluation comments, what we see is that the measure score for non-dual episodes is around 1.04. And that's also the area that we see for overall measure score as across all episodes.

And so that suggests that the dual and non-dual is actually quite close. At least in

that segment of hospitals where duals are very heavily concentrated.

CO-CHAIR DAMBERG: So, I'm mindful of time But, I think what I've heard is probably the socioeconomic demographic adjustment conversation is not at an end. And we encourage you to continue to work to evaluate the impacts of SES factors on this measure.

Any other comments before we move to voting?

(No response)

CO-CHAIR DAMBERG: All right. Seeing none, let's tee up the voting.

MR. SINGH: Okay. Well, thank you Cheryl. So we're going to begin voting on validity for Measure Number 2158. Your options are one, high; two, moderate; three, low; and four, insufficient.

Please vote now.

(Voting)

MS. O'ROURKE: If you could submit your vote through the comment. Or you could just

say it over the line too if you're comfortable doing that.

MR. SINGH: We're about to break -- we're about to break our record because we're at 20 right now.

MS. O'ROURKE: Okay. So we'll have to reset again. Sorry about that folks.

MR. SINGH: So, I'm going to reset the voting. So please bear with us one second.

MS. O'ROURKE: Everyone else on the phone, we have your votes. You don't need to send them again.

Larry, we do need yours still though. And Dr. Weintraub as well too.

MEMBER BECKER: This is Larry. I got disconnected. Did you get my vote?

MS. O'ROURKE: No, Larry. So if you could submit it again that would be appreciated. And we need to revote in here too. So, you can submit it when you get a chance.

MR. SINGH: Okay. So for those that are in the room right now, please submit your

voted.

(Voting)

MEMBER BECKER: Did it come across?

MS. O'ROURKE: Yep, Larry. I got it.

MEMBER BECKER: Great. Thanks.

MS. O'ROURKE: And Bill Weintraub, could you please enter your vote as well?

MR. SINGH: Okay. So all votes are in. And voting is now closed for Measure Number 2058 with regards to validity.

The results are as follows: three votes for high, nine votes for moderate, five votes for low, and zero votes for insufficient. With regards to percentages, that is 18 percent for high, 53 percent for moderate, 29 percent for low, and zero percent for insufficient.

This measure passes for validity.

MS. O'ROURKE: Just to be fair. We need 60 percent, Elisa, Helen? Or at 53 are we at consensus not reached?

DR. BURSTIN: High or moderate.

MS. O'ROURKE: Oh for -- sorry.

Apologies.

CO-CHAIR DAMBERG: All right. So, we are now going to move to feasibility. And John, I think you appoint to -- just a quick overview on that.

MEMBER RATLIFF: So feasibility we should be able to move through quite swiftly. The staff recommendation was high feasibility.

The data elements that are utilized for this measure are routinely generated already by the developer. And routinely used in day to day practice.

There won't be any kind of increased demands on practitioners. Or any kind of increased demands on the developer for use of this measure.

One of the comments that was brought up in the opening comment period from the committee was that while these measures are routinely developed, and it will be feasible for CMS to do this, it will be difficult if not impossible for independent calculation of the

measure.

Without presumably going through RSDAC or going through another licensing agent for CMS data and then obtaining that. So while this is feasible for use by CMS, it may be difficult to independently verify CMS' calculations or independently calculate using the metric.

Otherwise, no one else had any concerns with feasibility. And I would agree that it's a feasible measure.

CO-CHAIR DAMBERG: All right. Any comments?

(No response)

CO-CHAIR DAMBERG: I'm seeing none in the room. Anyone on the phone?

(No response)

CO-CHAIR DAMBERG: All right. I think we can move to vote.

MR. SINGH: All right. Thank you, Cheryl. So we're going to be voting on feasibility for Measure Number 2158. Your options are one, high; two, moderate; three, low;

and four, insufficient.

Please vote now.

(Voting)

MR. SINGH: All right. We were so close. But we're at 19. Which is over our denominator.

So we're going to have to restart again one more time. So please bear with me while I reset the slide.

All right. So for those that are in the room, please resubmit your vote.

(Voting)

MS. O'ROURKE: And is Bill Weintraub, are you on the line still?

MEMBER RATLIFF: Perhaps you're on mute?

MS. O'ROURKE: Bill, is that you on the line?

MR. SINGH: I guess all votes are in. And voting is now closed for feasibility on Measure Number 2158.

With the results is as follows: 12

votes for high, five votes for moderate, zero votes for low, and zero votes for insufficient. And that gives us 71 percent of the votes for high, 29 percent of the votes for moderate, zero percent of the votes for low, and zero percent of the votes for insufficient.

And this measure passes for feasibility.

CO-CHAIR DAMBERG: All right. Thank you. So, our last criteria, our lead discussant had to leave early. So, I'm going to throw it open to the committee to see if anyone has any comments or issues related to the usability.

And let's start with Brent.

CO-CHAIR ASPLIN: So, the first comment should not affect our voting today. But, this just for CMS in the room.

You know, I don't think it's appropriate for this measure to be used for physician accountability in MIPS or a value modifier program. It wasn't specified or built for physicians.

But like I said, that's a quick comment. And I don't think it really affects our task today.

Two comments or suggestions for the usability of the reports. One is state and national averages are helpful. It would be great to do a column on hospital referral region. Which I don't think would be that difficult.

But, there's a lot of variation across states. And to some of the comments around rural/urban, et cetera, an HRR column in addition to the state and national, I think would be more informative in terms of your local marketplace.

And then the other comment is that the subcategories of spend are really what's important. It would be really great if you could, kind of similar to the HealthPartners total cost index and the resource use index, if you could do utilization percentages within major diagnostic classifications for the various levels of post-acute care.

Because there may be good bench

marking data out there. But I'm frankly not aware of good bench marking data to drive performance.

And next site of care is the most important decision in total episode spent for people coming out of an anchor admission. And by diagnostic category would be really great to help guide clinicians around.

You know, 20 percent of beneficiaries with diagnosis X or DRG X for this category need skilled nursing care. And your average is 35 percent for example.

And the problem with the reports as they are today, is it's just dollars spent. Which is also helpful.

But it isn't granular enough to really drive action or improvement in understanding where you are relative to at least average utilization for the various levels of care. Does that make sense?

CO-CHAIR DAMBERG: Nancy?

MEMBER GARRETT: So, I also have, I

guess suggestions on how to make this more usable. So, I think one of the reviewers commented that in practice this measure is difficult for hospitals to consume because the data is not readily there to really understand what's driving their score.

And I do appreciate the files that you make available. What we found is that it takes a lot of effort to get those files and put them together in a way that gets you some answers.

And even then the answers aren't complete, because all the data elements aren't there. So you can see for example, what post-acute providers are referring to most. But not Prado it by where the biggest costs are.

And so just if there's a way to make that system more usable, I think that would really help the ability for people to -- for providers to act on this.

As it is, it's kind of another data point that we know we're being measured on. But it's really hard to take that next step of, okay,

what are we going to do to improve? So.

CO-CHAIR DAMBERG: Thanks Nancy. Sri?

MEMBER SRIDHARA: Yes. So I'm probably following up with somewhat -- just some more comment. So it feels like a very helpful, quick first measure.

That I think when we do our work we usually try to implement it so that it's a good measure you can get out the gates with. But it doesn't help you say, what should be your action steps next? So, it feels like a bit of a blunt instrument is what it amounts to.

And so what we usually do is put this in and then have some other episode measures to where we can do some of this more refined parsing of data to drive action. So, that's really it.

CO-CHAIR DAMBERG: Great. I'm not seeing any more hands in the room or online. So why don't we move to vote on this. Because I know we have a few more agenda items to go before the end of today.

MR. SINGH: All right. So we're going

to begin voting on usability and use for Measure Number 2158. Your options are one, high; two, moderate; three, low; and four, insufficient.

Please cast your vote.

(Voting)

MR. SINGH: We've gone over our denominator this time. It's like a pattern here. So, I'm going to reset the slide. So bear with me for one minute.

MS. O'ROURKE: And just for the record, our denominator is going to be 18 for this count. Because Bill did email his votes to us.

So, just FYI if you notice a change in the denominator.

MR. SINGH: Okay. So if you'll please submit your vote again. Those that are in the room.

(Voting)

MR. SINGH: Okay. So all the votes are in. And voting is now closed for usability and use for Measure Number 2158.

And the results is as follows: five votes for high, ten votes for moderate, three votes for low, and zero votes for insufficient. That gives us 28 percent of the votes for high, 56 percent of the votes for moderate, 17 percent of the votes for low, and zero percent of the votes for insufficient.

This measure passes for usability and use.

CO-CHAIR DAMBERG: Okay. One more vote to go. This is the overall.

MR. SINGH: Yes. So we're going to be voting on the overall suitability for endorsements for Measure Number 2158. Your option is one yes, and two no.

So please cast your vote.

(Voting)

MR. SINGH: For those that are in the room, can you please submit your vote one last time. We're missing one.

(Voting)

MR. SINGH: Okay. So all votes are

in. And voting is now closed. For overall suitability for endorsement for Measure Number 2158, the results is as follows: It's 16 votes for yes, and zero votes for no.

Oh, so -- well 16 votes for yes and one vote for no. That gives us about 98 percent --

MS. O'ROURKE: Yes.

MR. SINGH: So 98 percent for yes.

MS. O'ROURKE: Could we redo that just for the record? Just to be safe. People on the phone, we have all your votes. So, in the room, please?

Hopefully this is our last vote of the day. And we'll try to find the source of the gremlins or Russian hackers, whatever you want to say they are.

MR. SINGH: Got some at NQF -- there is -- okay. So for those that are in the room, please resubmit your vote.

(Voting)

MR. SINGH: So the last one with

feeling this time. So, for overall suitability for endorsement for Measure Number 2158 is as follows: 17 votes for yes, and one vote for no.

And that gives us 94 percent of the votes for yes, and 6 percent of the votes for no.

CO-CHAIR DAMBERG: All right. Thank you. I just wanted to quickly see if there was any additional member or public comment at this time, both in the room.

And Operator, if you want to ask if anyone on the phone has a public comment.

OPERATOR: Okay. At this time if you would like to make a comment, please press star then the number one.

(No response)

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

CO-CHAIR DAMBERG: Great. And there are no comments in the room either.

We are going to take a very quick break. And I would ask people if they could be back in the room by 2:30. Because I know the NQF

staff have a few more items they want to get us through.

But, they also know we need a leg stretch. So, back in a few minutes.

(Whereupon, the above-entitled matter went off the record at 2:23 p.m. and resumed at 2:33 p.m.)

CO-CHAIR ASPLIN: Good afternoon. What we have left this afternoon is an overview of the cost and resource use measurement landscape. And then, some discussion about episode-grouper based measures. And the overall project update for cost and resource use. And the staff are going to lead us through this conversation, so I will turn it over to Erin.

MS. O'ROURKE: Thank you. And I want to introduce Ashlie Wilbon, who has joined us at the table. I think she's worked with many of you in the past. So, Taroon, Ashlie, and I can kind of drive the slides and we're going to skip around a little, so feel free to jump in if you have any comments.

So, if we could skip to Slide 28, let's keep going, because Rachel went through this in the morning. Again, you saw this slide. So, again -- sorry, one back. We did just want to quickly refresh you on what measures are currently in the cost and resource portfolio.

We covered three of these today, the others are the ones that we actually just looked at over the summer, the episode of care for AMI, heart failure, and pneumonia. Next slide.

So, to just give you an idea of where these are used in federal quality initiatives, as the team just was saying, the spending preventing measure is used in the Hospital IQR program, as well as the Hospital Value-Based Purchasing Program. You can also see similar use of the three episode of care measures, with the exception of the pneumonia measure is right now only for reporting. Next slide.

So, we're going to kind of combine the next section, if that's okay, and then, open it back up for any input the Committee might have on

gaps. So, I wanted to just briefly give you an idea of some of the measures that are coming through the pipeline. Next slide.

These are some that we've seen through the MAP process, as well as in recent legislation. Just to give you a quick refresher on the MAP, the MAP is the body that NQF convenes to provide input on the selection of measures for federal value-based purchasing and public reporting programs.

Each year, the MAP is given a list of measures by HHS that they're considering implementing through the rule-making process and does a review, going measure by measure and making a recommendation about whether they support its use or not.

Similarly, MAP also does some work to identify gaps in measure development and does work to support alignment across the public and private sectors, as well as settings, different levels of analysis and populations to try to promote care coordination and reduce the data

collection burden. Next slide.

We've been working to try to integrate the MAP and CDP processes more closely. We know that endorsement and selection are really two processes that need to work in parallel and we want to make sure that you're getting the information that you need from the MAP work to support your conversations about endorsing the measures.

Similarly, we want to make sure we're giving the MAP what the CDP standing committees feel is the most relevant information about the measures that are coming under consideration for the various federal programs.

So, staff has started to undertake some work where we will reach out to measure developers who have measures under consideration for MAP to encourage them to bring them in for endorsement work.

We also, you may have noticed in the preliminary analysis, include some information about MAP's review of measures in the use and

usability section. Similarly, when we do a preliminary analysis for the MAP committees, we include relevant feedback from the CDP standing committees and your reviews on any measures that have come before a standing committee before they go to MAP.

So, we do want to make sure that information is flowing between the two processes and both our endorsement and selection committees feel supported by the information we're providing. Next slide.

DR. AMIN: Erin, before you move on --

MS. O'ROURKE: Yes?

DR. AMIN: -- could we go back to that slide for a second? So, I just wanted to reiterate for the group, many of you have been part of the standing committee since we originally started our work in cost and resource use measurement, and so, I wanted to just take a second to reflect on the fact that we went from having -- actually many of the measures that we looked at today, these were the first measures

that we ever looked at. And so, now, three years later, we have a significant number of cost and resource use measures that are endorsed.

So, as Erin talked through, in terms of the current measures that are in the portfolio, what's really important for us is to get some guidance, and we can get it today and we can certainly get it from you all following the meeting today, around where we want this domain of measurement to go, in terms of, we sort of describe it as gaps, but we've talked about, actionability has come up a few times during our conversation today, the use of these measures in different settings, condition-specific measures, but we also want to have a coherent portfolio. This is an opportunity for an area of measurement where we can be much more strategic in the types of measures that we want to see.

Secondly, and as Erin described and I'll just highlight, we have the opportunity in the fact that this is a new area of measurement, but also a really important area for use of these

measures in programs, where we have the opportunity to see these measures, proposed measures, early in the MAP process, sometimes before they even get to the standing committee.

So, what Erin's going to walk through is some of what we believe is going to be coming down the pipeline for you to be evaluating in the next several cycles of our work. And I say several cycles, because it looks like there's going to be quite a bit of work coming down in this area of measurement.

But we want to make sure that you all are familiar with the fact that, as we have these conversations, the purpose of our discussion is not only for our reports and for transparency purposes, but it also feeds the MAP process as the MAP committees think about using these measures in programs.

And so, the preliminary analysis and the scientific review that's done in these committees have multiple uses and is really the foundation for the MAP recommendations for the

various programs.

And so, we look at these two processes as interconnected and the flow of information going from one process to another, with the staff really serving as that conduit. And so, with that, I'll turn it back over to Erin, as we talk through some of the measures that have been proposed in the various MAP work groups.

MS. O'ROURKE: Sure. So, on that note, next slide please. To just give you an update of what we've seen through the MAP process that has not really come before the Standing Committee since you've last convened, in 2014-2015, MAP took a look at a number of ACO-level measures for the Medicare Shared Savings Program.

One is an application of 2158, the Spending Per Beneficiary Measure you all reviewed today, however, at the ACO level, not for the hospital level of analysis as you reviewed today. Also, the Total Per Capita Cost Measure that I believe this Committee evaluated a few years ago and was not endorsed. Similarly, MAP did not

support that measure for implementation in the Shared Savings Program. Next slide.

So, this, I apologize for, the font is rather small, but in the year before last, MAP actually looked at quite a few measures under consideration related to cost and resource use. You can see there are a number of episode-based payment measures for various conditions that were under consideration for some of the hospital programs, the IQR and the value-based purchasing program.

There was also a spending per beneficiary of post-acute care for the home health, inpatient rehab facility, long-term care hospital, and skilled nursing facility settings. Next slide.

DR. AMIN: So, Erin -- sorry to keep jumping in.

MS. O'ROURKE: No, no.

DR. AMIN: Can we go back to this slide? So, one of the things that we wanted to point out with this is, there is a real need,

again, just going back to my previous comments, there is a need, as you can see with the number of measures that are coming out for cost and resource use and what they're focused on, there really is a need for guidance around how and which types of measures we're actually developing for cost of care measures.

And you can see the myriad of conditions that have been already proposed, and they're all -- essentially, the recommendations to the MAP were based on the question of whether these are important areas for cost measurement. They might be important clinical areas, but are they really important cost drivers for the system?

And so, again, we may need to really think about, as part of our work, as we think about the next phases of work, how do we really think about breaking down these episode-based measures and how do we come up with a conceptual model for how we're going to be able to prioritize?

And give some guidance to the field about what types of measures we want to be seeing here for cost of care, because some of these are already being developed, there's a lot of development dollars going into this, we want to make sure we're really addressing some of the high impact areas and in the last row, high impact care settings, and making sure that we have a level of harmonization of the measures that we're seeing across these various settings.

MS. O'ROURKE: So, this is -- actually, to go back to a point, Brent, you raised during one of our orientation calls was, legislatively, there have been quite a few updates around cost and resource use measurement.

In particular, the Improving Medicare Post-Acute Care Transformation Act of 2014. The IMPACT Act required the development and reporting of measures of resource use for LTACHs or SNFs and home health agencies.

And then, the Medicare Access and CHIP Reauthorization Act of 2015, MACRA, which

repealed the sustainable growth formula and streamlined clinician quality reporting programs and incentives into the Advanced Alternative Payment Models requires the use of cost and resource use measures. Those are included in both tracks of the Quality Payment Program, both the new Merit-based Incentive Program, as well as the Advanced Alternative Payment Models.

So, just to give a very high level overview of the MIPS Program, MIPS includes four domains: quality, resource use, clinical practice improvement activities, and advancing care information.

The resource use domain will include a number of measures, including the Total Per Capita Cost for All Attributed Beneficiaries, Medicare Spending Per Beneficiary, and up to other 41 episode-based measures. These are still in development, but focusing on acute care, procedural, and episodic care. Next slide.

So, I want to kind of reframe some of these questions here, pulling in some of Taroon's

points about not just what information the MAP needs, that you can see on these slides, but also to go back to the question about, does the Committee have input on what are some outstanding gaps in development?

Any guidance you'd want to put there of where we should try to perhaps make recommendations around where to put development dollars, any outstanding gaps you see in the portfolio. So, in combing the past two sections, on both the MAP work and any gaps that remain in the portfolio.

DR. AMIN: Recognizing that many of these forty-some-odd episode measures may already be in development.

MS. O'ROURKE: Yes.

DR. AMIN: So, some of that may already be in development. So, I think, really, the fundamental question here is, as we think about these two important processes, are there any sort of recommendations, observations that you have as the Committee around how we can continue to

improve the information flow and also, where we might, as a Committee, focus on prioritization of guidance to the field for this area of measurement?

If we're looking at doubling, potentially even tripling the size of this cost of care portfolio in the next three to four years, how do we start to rationalize what we really need to be looking at in this cost of care domain?

CO-CHAIR ASPLIN: So, with that, we'll open it up for discussion and comments from Committee Members. Srinivas, do you have a comment?

MEMBER SRIDHARA: Still up from --

CO-CHAIR ASPLIN: Cheryl?

CO-CHAIR DAMBERG: So, it seemed as though, from that first slide that had all those different conditions, there's pretty heavy emphasis on the hospital setting. And I guess I just need to understand better what's in the measure development pipeline vis-a-vis cost for

ambulatory care and ambulatory care conditions.

MS. O'ROURKE: So, I think that is something we haven't really seen yet, other than the language in the proposed rules about the up to 41 measures. I was just looking today, there's some commenting open on the MACRA website, with some information about the various episode groups and the triggers.

But those measures are still in development and we haven't really seen what they might look like, they haven't come to the MAP process yet.

CO-CHAIR ASPLIN: Jack?

MEMBER NEEDLEMAN: We spent a fair amount of time talking about usability today, even as we abbreviated it in some cases, but I think that's critical. And I don't think the way we're getting the data on these measures, either we or the MAP, is facilitating our assessment of the usability or the appropriateness of the use, given the validity of the measure.

An awful lot of the payment measures

wind up in a grid system, with very sharp edges, based upon your location in a distribution, not the absolute level of your spending or resource use.

And we have not been asking -- we got told how many people changed quintiles, but we didn't get told how many people, how many hospitals or how many systems or how many provider groups wound up being shifted from one payment category to another, either a bonus to nothing or to a penalty.

And yet, those are the way the measures are being used in the systems that the MAP is evaluating. And the assessment of how well they perform needs to be done against use, not more abstract levels of r-squares or anything else.

So, I think that's one of the key lessons I draw from this, for both our group, our work, and the MAP work, in terms of enhancements and guidance on how to do it and guidance for the developers on what to provide.

I think one of the issues for us, we are the Cost and Resource Use Committee, we often don't measure resources because of standardized pricing, the issue has been, in the value-based payment, was the fact that people were looking at value.

And I guess, one of the questions, in terms of where the appropriate forum is, is, are these grids, which integrate both information on quality and information on resource use, a measure in and of themselves or is it only the separate components that the measures and, therefore, how do they get evaluated?

And to the extent that the grid is the measure, should the grid be coming into this Committee with a broader mission of assessing cost, resource, and value measures? So, those are the issues I would raise for consideration.

CO-CHAIR ASPLIN: Martin?

MEMBER MARCINIAK: So, four years ago, we came here and we had basically the same conversation we had today and I made the comment

to a couple of you in passing as we went through today, which caused me to kind of scratch my head a little bit.

To Janis's point, the dialogue this time seemed much the same, but the vote was very different. Four years ago, I found, mistakenly, in actually revoting at some point to sort out where our true feelings were about the measures.

But at the end, we end up with a conversation of, well, what are we measuring? We're not measuring cost, we're not measuring resource, and, thereby, we're not measuring value.

And so, a couple of years ago, I got excited about the fact that I actually started seeing measures coming through that were what I would call therapeutically-aligned, so, the pneumonia one, right?

So, how can you measure cost without context, was the conversation four years ago? Once again, today we had a conversation of cost without context, and that's the part that I have

a hard time resolving when I go back to the office is, well, what do you share with people?

Well, we passed three resource measures, that's what we're calling them. They don't measure efficiency and they don't measure value. And so, I would like to see, as we continue to move forward with these conversation, a degree of trying to drive the dialogue together a bit more.

Because, at least at that point, it will become more meaningful, because then you'll understand what's happening in an ambulatory setting, you'll understanding what's happening in a hospital.

And, frankly, to the comment that somebody had made about the group that was here from Yale three years ago that got a really hard time, and then, the young lady gave me a hard time, because I explained that I couldn't understand what she was doing well enough to tell my parents, who use these systems.

There seemed to be, again, a lack of

congruency with respect to what the ultimate pull-through ends up being. So, we approve something here, we're not always sure exactly how you differentiate what's the patient populations, problematic, and then we approve it and it's used broadly and we don't wrestle with that which we have wrought, really.

So, that's my two cents for this. I would like to see more communication, not less. So, you tend to wax and wane based on the level of activity, but there were some things that were on the, I guess the chart, that I don't recall having seen. Things that we had elected not to recommend out to the broader group. Some of them, again, were disease-specific, sort of resource and cost total.

It would be nice to see a little bit more of that, because that would allow me to reestablish in myself some context of, that we are trying to link the stories between cost, resource efficiency, and value. Because most of my day job is about talking to hospitals and

other groups about what value is and I still haven't found a good definition of what that means.

CO-CHAIR ASPLIN: Troy, and then, Dolores.

MEMBER FIESINGER: Hello, this is Troy. Appreciate hearing long-term care mentioned, because the question for me and my practice and being a family doctor is measures that look at long-term treatment in the outpatient setting, diabetes, heart failure, et cetera.

The sub-population that comes up very often in our discussions with management of care plans especially is, those high-cost, high-risk acute patients.

And I don't know if there are any measures looking specifically at that population that have multiple illnesses, are in the hospital often. Are there any benchmarks to compare how we're doing with them? Do they just cost a lot and that's the best we can do, or are there areas to improve?

Second, in the post-acute care setting, that are ACO, really looking closely at SNF use versus LTACH use, home health PT versus going to a clinic for PT, so care setting measures addressing those subsets of use for those services, but not just in the post or discharge setting, but home health in particular.

A pattern I see is home health being used and really -- care, but that's not covered. I appreciate the comments that we just heard on value with the care bill especially increased risk for lower cost, but all we ask, do a good job quality-wise. Measures that balance those two with the value would be very helpful.

CO-CHAIR ASPLIN: Thank you. Dolores?

MEMBER YANAGIHARA: So, my comment is related to feedback that we get from providers all the time, in wanting to have a consistent set of measures that they can be measured on across their whole patient population.

And I know that there is lots of reasons why that's challenging, because not every

measure is applicable to every population, but if you think about it, the HealthPartners measures were specific to commercial, and then, there is the Medicare measure we just looked at, was specific to Medicare.

And there may be reasons for that, but it makes -- it drives the providers crazy to have similar measures, and those two aren't exactly similar, but that only apply to part of their patient population.

So, I mean, I don't -- like I said, there's lots of reasons why that gets very challenging to do, but -- and then, that also has to be balanced with, sometimes we start getting -- so, we have total cost of care is not as actionable, but it's applicable to everybody. And then, you start drilling down into very condition-specific thing that are maybe more actionable, but then don't apply to as large of a population.

So, like I said, I recognize there's a lot of challenges, but I think to just keep in

mind that the more we do kind of payer-specific kind of stuff, the less impactful it will be, because the providers are trying to respond to all these different signals.

CO-CHAIR ASPLIN: Larry?

MEMBER BECKER: Yes, thank you very much. So, I wanted to reflect on a couple of the prior comments and particularly Jack's. And perhaps this is what Jack was saying, in that there are a lot of measures and there are a lot of measures outside of the cost and resource use measures.

And understanding those may provide context and a larger picture and fill in some of the voids in maybe our perceptions, maybe our realities about what we're looking at on cost and resource use.

And that it would be helpful, perhaps, to see where there may be complementary measures that are sort of filling in those blanks, because I have a sense that a lot of these measures are aimed at trying to get a complete picture and

when we look at it from simply the cost and resource use perspective, we're sort of getting a narrow context, if you will.

And so, to the extent that other things that are happening to fill in measures could be complementary, I think it would be helpful for us to at least see and understand those in the context of the work you're asking us to do.

CO-CHAIR ASPLIN: Thank you. Jim?

MEMBER NAESSENS: Thank you. Actually, I was going to reflect back on more what Martin said and also Dolores. And I was going to give almost the opposite recommendation that Dolores gave, that as measures proceed and as we're kind of evolving down the path, and we look at the usability for the providers or we look at the usefulness for the consumer, global measures are less useful.

They don't give us an idea of where we can change, what's the best one. I'm coming in with cancer, I really -- maybe the cancer

practice here is very different than all the other practices.

So, the more and more we can move towards measures, maybe not disease-specific or certainly not payer-specific, but if we think of a clinical domain, whether it's obstetrics, whether it's primary care, whether it's orthopedic surgery or some groups like this, and recommend that measures be focused in those fairly extensive, large domains, we should be able to bring more of the quality measures with the cost and utilization measures, get more meaningful measures that are useful for providers to make changes and also useful for consumers to start selecting places.

CO-CHAIR ASPLIN: Srinivas?

MEMBER SRIDHARA: Yes, I think Jim mostly said what -- similar comments. So, thank you. If I could extend that, I think in terms of a role for this group, in particular, is I think the process being described, a lot of what's coming and sort of how can we provide guidance,

and it almost seems like taking this conversation and providing a framework to what is needed and crafting that is what it's necessary.

Because it's -- to also go back to something Martin was saying, in terms of aligning with the therapeutic context, or if you think about for, whether it's a hospital, a clinician, anyone, they're operating in specific service lines or categories of service that they're doing in general, whatever that is, or specialities, and we need something that aligns well with that, so that it drives that action.

It could be that you still have measures that are applicable across, like, these cost of care measures, like the total cost of care measure we looked at today, but there's probably a push to say, in the coming time, we want those to be applied in specific service areas and contexts and bring that through, and that's how we would like to approve it.

And that might be a way that, even for CMS or others that are going to be using this in

how they, say, do MIPS or something else, then they have a sense of how to tailor it to the different service lines. But I think we could probably help to provide that framework as a group.

CO-CHAIR ASPLIN: Thank you. Kristine?

MEMBER MARTIN ANDERSON: I think when -- so, what I struggle with is, whenever we have these measures that have layers and layers of algorithms, right, so they're going to have a grouping algorithm, they're going to have a measurement algorithm, they're going to have a risk adjustment algorithm, et cetera, that it's never true that it works equally well for all populations.

And so, one thing that I struggled with today and I think on this question of use, that we need to continue to struggle with, is that, if they don't work well for a particular population, so a population-based measure does not work well for individual physicians or certain types of characteristics of individual

physicians or some places with a claims stream really does not well represent the care process, there isn't a way to articulate, here's where it shouldn't be used.

And so, once they get endorsed, we have these conversations, but once they get endorsed, they can get used anywhere. And so, I know that NQF has been hesitant to walk into the -- down the road of appropriate for what use, but when you get to these types of measures that have so many layers of algorithms and they're reliant on blunt information at the base, I think it's important that that kind of conversation be had in the endorsement process or the MAP process or somewhere.

CO-CHAIR ASPLIN: Thank you. Nancy?

MEMBER GARRETT: So, one thing that I've been thinking about in regards to this portfolio is just about price. And so, whenever I've been involved in total cost of care type measurements across populations, I'm always struck by how much price is a driver of the

results. Utilization is important too, but price is really big, in terms of differentiating performance.

And what role would NQF have there, in terms of having more measures of price and having more price transparency? So, that's kind of a political question, but an important one, I think, because it's a really key part of cost.

And there is some work in Minnesota, where Community Measurement has created some, really, price measures for several procedures that are high volume, like MRIs, where we've aggregated average allowed price across all payers, so that for one -- you can compare the cost of that procedure across all providers.

And that's on our website. That's something we've done as a community, but it's not a perfect measure by any stretch, but it is kind of a baby step towards more transparency. So, just something to think about.

CO-CHAIR ASPLIN: Thank you. I'm going to make a quick comment and then, I'm going to

hand it over to Ashlie and we can move into some of the grouper conversations, if that's all right?

To me, the urgent area is around physician accountability around resource use, because we don't have any endorsed measures there, and MIPS is coming up. And I just think this whole process is way too labor intensive to use the CDP to try to sort out MIPS.

And I don't know what level of stakeholder engagement and dialogue there is across CMS and MAP and physician groups, et cetera, around, what is this resource category of MIPS really going to look like, and Advanced Alternative Payment Models?

We need a total per capita cost measure in Medicare, we need one. I know it didn't get over the bar when we talked about it last time, the total per capita spending, but I don't see primary care groups and physicians to be accountable for episodes, it just doesn't make sense.

And so, there needs to be some version of the HealthPartners measures or something like that in Medicare, I think. And where it fell apart was on attribution, that's where it fell apart.

And I think we have to sort through that by having dialogue about use, which is not part of the CDP, the Consensus Development Process, but if CMS could clarify use and there could be at least a modest amount of agreement among physician groups and CMS about how the resource use category of MIPS is going to move forward, so that you could Pareto chart out all the docs and this is generally who we're going to use a total per capita measure for and these are the physicians we're generally going to use an episode-grouper for, and then some ambulatory non-procedural specialists who are in single speciality TINs are going to be left out.

And I don't know what to do about them, right? So, what do you do about rheumatology? Huge spend, not really going to

fit a grouper, and if you're a single specialty rheumatologist group or solo practitioner, where do you fit?

I don't know, but let's solve for the 20 after we get the 80 figure out, right? I don't have the answer for everything, but I don't want to go through 45 measures in this process and then find out that that's not how they're going to be used. Helen?

DR. BURSTIN: Yes. I think so many of you raised great points and lots of work I think we all need to collectively do for the future. We do know that the measure we just talked about, the MSPB Measure, is being respecified to the physician-level, to the TIN-level, so that will be coming back to you. And they'll adapt -- again, it may not answer all of that, but at least that is one thing we know that is coming down the pipe.

But, again, you've raised a lot of really important issues about really thinking about what's needed, frameworks, et cetera, that

I think are exactly what we'd like to have come out of this incredibly smart Committee.

DR. AMIN: Yes. So, I think these are excellent comments. I think there's a lot of ideas here, in terms of -- Ashlie, did you have some comments? I didn't mean to jump in.

MS. WILBON: No.

DR. AMIN: We're good. So, I think there is a lot of work that we can do in terms of just even thinking about how we start linking these cost and quality measures, whether that's a measure developer responsibility, as the Linking Cost and Quality Expert Panel sort of recommended, or based on Srinivas's sort of proposal. That might be some activity that this Committee can do, if we propose some additional funding to be able to support that type of work.

So, Erin, if it's okay, we can move on from this section, and I'll sort of move on to the next slide. And this is a little bit of a shift of gears and I appreciate everybody sort of working with us.

And we have a few different conceptual things that we wanted to talk about and we're trying to stuff it all in the last hour here. So, we're going to shift between multiple different topics.

So, the next discussion, you have a handout that's titled, it's on NQF letterhead, that's titled Cost and Resource Use Measure Evaluation Criteria Update Recommendations. So, the purpose of this document is to review the CRU criteria that are used for endorsement.

And over the past few years of experience we've gained from deploying these criteria in multiple CDP projects, there have also been updates to the quality measure evaluation criteria as well.

And so, the purpose of this document in front of you is to -- there are several recommendations that staff is making to update the criteria for the Standing Committee to consider. And from our perspective, these are intended to simplify the criteria and make them

more precise.

In terms of, there are components that are included here that are recommended for removal, that we don't argue are important, however, it is very unlikely that the Standing Committee has ever made recommendations to not endorse a measure based on these criteria.

And so, for us, they don't feel like criteria if we're not really using them to distinguish between different measures. So, the first is on Criteria 1A, the measure focus. Currently --

MEMBER WEINTRAUB: Bill, just a second. For those on the phone, which one of the documents is this?

DR. ROILAND: It should be the one titled Cost and Resource Use Measure Evaluation Criteria Updates. If you don't have it, we'll send it out to you right now. Apologies about that, if you can't access it.

MEMBER WEINTRAUB: I don't see that. I'm looking one the website, I don't see that.

So, if you'll email it to me, I'd appreciate it.

DR. ROILAND: Okay.

CO-CHAIR ASPLIN: Maybe, can we also bring it up on the screen for the webinar? Would that be too much to ask?

CO-CHAIR DAMBERG: You can also post it to the webinar.

MEMBER WEINTRAUB: That is too much to ask? Okay.

DR. ROILAND: Yes, could we screen-share this? Do we have it?

CO-CHAIR ASPLIN: Can we vote on it?

(Laughter.)

DR. AMIN: That was good Brent, that was good. Sorry, Bill, we're pulling it up and then, we'll also send it to you.

MEMBER WEINTRAUB: Great.

DR. AMIN: And I'll do my best to actually walk through it in detail. So, the first criteria, Criteria 1A includes two components.

The first is that the measure focus

address a specific national health goal/priority identified by HHS or the National Priorities Partnership convened by the NQF. And we are recommending removal of this component of 1A, since, really, measuring costs conceptually is very difficult to argue is not important to measure.

And, really, it's the second half of this criteria which is related to what we're actually looking at, actually demonstrating that it's a high impact aspect of healthcare, meaning that it affects a large number of individuals, that it's a condition, potentially, that's affecting a large number of individuals, or a leading cause of morbidity or mortality, or it's just high resource use and severity of illness.

And so, this is really the component that is of importance here. And so, this is where we recommend that we continue to maintain focus and remove the first component. The second is around performance gap. Clearly, we've had a lot of discussions around that today, I think

that stands.

And then, the third is, 1C, the intent of the resource measure and the measure construct are clearly described and the resource use service categories that are included in the measure are consistent with and representative of the intent of the measure.

And so, the challenge here is that, oftentimes with this section of the evaluation criteria, is that, the importance to measure and report criteria is intended to be conceptual in terms of what the intent of the measure is supposed to be focused on, whereas scientific acceptability really gets us into the measure specifications.

And so, 1C often has the tendency of drawing us right into the most -- the heart of the -- where people are really concerned about the specifications, where really that's not the intent.

And, quite frankly, it is repetitive of what we're looking at with the reliability,

meaning the specifications are precise, and then, the validity, meaning the specifications are consistent with the intent.

And so, we are recommending removal, since this is really repetitive and, quite frankly, would be more structured if we included it in the reliability and validity evaluation of the measures.

And then, for reliability, we just included, this is to be consistent with the quality measures evaluation criteria, which is looking at the specifications. We have not had an eMeasure specified cost and resource use measure yet, but it likely is coming. Cheryl might be working on it.

(Laughter.)

CO-CHAIR DAMBERG: It's going to be a hybrid.

DR. AMIN: Yes, or at least a hybrid. But either way, at least just making it clear that we're looking for ICD-10 specifications and then, we're using the eMeasure specifications,

HQMF format.

And so, this is not really -- this is just to be consistent with the rest of the quality measure evaluation. And so, that really is the bulk of what we wanted to bring in front of you today.

We believe that these are relatively minor recommendations, in terms of clarity and just to make your work actually a lot simpler, in terms of structuring our conversations. So, those are the recommendations that are in front of you for evaluation. Erin, is there anything else that you wanted to add, as it relates to this --

MS. O'ROURKE: No, I don't --

DR. AMIN: -- discussion?

MS. O'ROURKE: I don't think so. You covered it.

DR. AMIN: And so, I would welcome any discussion. And if there isn't any, that is okay, we'll take that as agreement and then move on. But I certainly welcome any conversation on

it.

MEMBER WEINTRAUB: This is Bill. To me, this just does add clarity.

DR. AMIN: Thanks, Bill.

CO-CHAIR DAMBERG: Nancy?

MEMBER GARRETT: So, I just found under 1B, the disparities in care across population groups, when we were going through that today, I found that kind of confusing.

Because people interpreted that as, should this be risk-adjusted for SDS? But then, that was in the validity section. So, I don't know if we're just not interpreting it the right way, but it just doesn't seem to fit there.

DR. AMIN: We struggle with that as well. The challenge is that this is -- yes. So, the challenge is that the intent of this in the quality measures evaluation criteria is that, one of the reasons for including a measure is that it might be sort of conceptually an area where there is disparities of care. Now, whether the measure is able to find the disparities is sort of the

validity question.

In our work, I don't know that it -- it is conceptually challenging to understand, because, again, we have that same problem, where we look at this and we're looking to see where there's disparities and is that a validity question?

That really is, that is a challenge as well and I think that it does offer some confusion as well. And it certainly added confusion in our preliminary analysis work, in terms of making sure that that was clear.

So, I mean, I would welcome additional recommendations on that. We would have to discuss with our measurement science colleagues on the quality side to understand whether that was an option to remove, but we certainly welcome thoughts on that, if you have any, Nancy.

MEMBER GARRETT: Do they have to be aligned to --

DR. AMIN: No, they don't --

MEMBER GARRETT: -- the cost and

quality measures?

DR. AMIN: They don't need to be.

MEMBER GARRETT: Okay.

DR. AMIN: They don't need to be.

DR. BURSTIN: It certainly would be optimal, otherwise we have to explain why they're different. So, if there's a reason to have something in there, there has to be a compelling reason why cost is different.

MS. WILBON: I wonder, too, if it would just be an issue of just clarifying language. I think this second bullet is really around, are there such disparities in the population that it's necessary to have this measure to help illuminate that there are those disparities?

So, I think that disparities in care -- exactly. So, I think it may be the wording of it and the way that it -- so, maybe we can come up with some new language, so that it's clearer when we're going through the evaluation that the context of the discussion is really around, is this measure needed to help us understanding why

there are disparities in care?

Conceptually, that this measure is needed, it's important to have this measure for that reason. So, I think there may be ways that we can massage that language a little bit.

CO-CHAIR DAMBERG: Any other comments? So, did you want to take a vote on this or, like, a show of hands?

(Laughter.)

DR. AMIN: I will just --

CO-CHAIR DAMBERG: No more electronic anything.

(Laughter.)

DR. AMIN: No, I think we're -- unless there is --

CO-CHAIR DAMBERG: Any objections?

DR. AMIN: -- any objections, we would just take it that way, I think it's fine. And then, I mean, this will have to go through other bodies to look at, since it is a criteria review, but we'll take that as input. Okay. So, Erin, I'm going to turn it over to you on the SDS

slides. Or whatever topic you want to cover.

MS. O'ROURKE: Why don't we do the SDS and then, I can turn it back to Ashlie for the groupers? And that gets us on-schedule. We've got attributions scheduled after the grouper work. So -- yes.

MS. WILBON: So, which topic are we on?

MS. O'ROURKE: I'm going to just quickly go through an update on our trial period for adjustment for socioeconomic and other demographic facts, or SDS as we've been referring to it.

So, just to refresh everyone on the last this Committee, really, heard of this, we convened you over the summer to provide input on the three measures from CMS, Yale, the episode-based ones.

When we received an appeal about the decision to inverse them without any SES factors in the risk adjustment model, the Committee recommended upholding endorsement of the measures as-is. That decision was ultimately ratified by

the CSAC and the Board upheld the endorsement.

So, we did just want to close the loop with the Committee on those. And thank you for all the extra work you did to weigh in on those. Next slide.

So, again, this is a topic that came up throughout the conversations today, so I don't really want to belabor it, but this is from the ASPE report that we discussed quite a bit already, just showing some of the challenges for getting this data.

And as Helen said, it's really been one of the major findings of our trial period is just how challenging it is to get the data to really link the empirical analyses to the conceptual basis that we're seeing.

And as Helen said, NQF is committed to continuing to do what we can to drive the better data. We've been recommending at the annual update that developers give us more information about where they are.

So, I wanted to just highlight to the

Committee, that's one of the significant findings, was this is still very challenging to do and we are committed to doing what we can to improve the underlying measurement science here. Next slide.

So, again, what we found has really been consistent with what's come out of ASPE. Medicare patients with social risk factors tend to have worse outcomes, regardless of provider. Dual enrollment in Medicare and Medicaid can be a powerful predictor of poor outcomes.

Providers disproportionately serving Medicare patients with social risk factors have worse performance on quality measures and experience somewhat higher penalties in value-based purchasing programs. And, really, tracking with what's come out of our trial period is better data and more study is needed here. Next slide.

So, some initial results of our two-year pilot. We really want to get the right data to assess social risk, what we've got currently

is just not sufficient to show up in the empirical analyses. We've probably only had ten measures or so of all that we've looked at over the past two years end up having SES factors in their risk adjustment models.

We've seen a lot more come through with a conceptual basis, but then it's really just been a challenge to have that hold up when you really crunch the numbers, similar to what, I think, this Committee grappled with this morning.

Again, we do want to improve the quality of care for social risk factors, this is a key challenge I think the system is grappling with. We need to find ways to better account for the impact of social risk, so that value-based purchasing programs reward providers fairly.

And we want to get to, really, the next generation of measures that are assessing healthcare equity. And this is something our Disparities Standing Committee is tasked with, they are actually meeting at the end of this month to really start to consider what measures

of equity could look like. Next slide.

So, some key upcoming milestones. The trial period will end in April 2017. During the June 2017 meeting, we'll reconvene the Disparities Standing Committee and ask them to evaluate the trial period and make a recommendation to the CSAC and the NQF Board about whether or not we should continue to allow SDS factors to be included in the risk adjustment models of endorsed measures.

So, is there any reasons to put the previous policy that prohibited standing committees from consider this back in place? But after the Disparities Committee makes their recommendation, it will go through the CSAC and the Board, I believe in July, for their input. Next slide. So, again --

DR. AMIN: So, Erin --

MS. O'ROURKE: -- looking -- oh, sorry.

DR. AMIN: So, I think, there's the Committee discussion around input to this Disparities Standing Committee, there's been a

lot of conversation that we've had over the past day related to this topic of SDS and risk adjustment.

I just wanted to sort of reiterate them as they relate to just input to the Disparities Committee, in terms of what I've heard so far. There's been a lot of discussion around unmeasured risk, that Helen described quite a bit in some of her comments, that really needs to be explored by the field, in terms of additional research.

Srinivas made the comment around, how do we think about data, particularly data for national measures, and what's actually available for these measure developers to really use, if we're expecting them to use for national measurement and payment purposes?

Janis made some points around provider and population-level adjustments, and what the approach is that we might want to consider in order to test the various measures that are in front of us.

The last point, I think, that has come up many times, both through our public comments and then, through comments around the room as well, is around the fact that, when we talk about risk adjustment, the complexity of the conversation is often very difficult for our stakeholder groups to keep up with, quite frankly.

The issues are so complex in terms of the data that we're talking about, the way the measures and variables are -- well, the way that the variables are constructed to represent the underlying construct, and then, the way that these risk adjustment models work in analyzing -- just even our conversation earlier around, what's in the ASPE report versus what measure developers are finding.

And actually having a relatively clear discussion in our reports and in our discussion around why there's these different findings. And, oftentimes, this complexity is really creating a lot of confusion and I think

frustration among our stakeholders around differences in what they're seeing in literature versus what we're finding as a Committee.

So, I think all of those are discussions that we've had today, in various forms. I would certainly welcome other feedback, but those are at least some of the key things that I've heard today for the Disparities Committee to consider.

MEMBER WEINTRAUB: I think something that could really be very helpful would be a white paper on the kinds of risk adjustment that NQF is looking to use. Maybe you already have it, but if you do, I haven't seen it.

And this can be written in a way that people can understand. Now, not laypeople, but that's not what we're really -- that's too heavy a lift. But I think that the consumers of these products can understand it, if it's well explained. And if you go and try to just gather that yourself from the literature, for most people, that's a really heavy lift.

CO-CHAIR DAMBERG: Larry, did you have a comment? I saw your hand up in the meeting.

MEMBER BECKER: Oh, sorry, no.

CO-CHAIR DAMBERG: Okay. Jack?

MEMBER NEEDLEMAN: I think this issue of SDS adjustment is a complex one, we've already figured that out. But I think one of the reasons is, we're trying to capture at least three different kinds of things with it.

And I believe that it would be helpful for committees considering this and for the developers to think about how well the way in which they're doing the risk adjustment or how well they're doing their analysis addresses the three concerns that we have?

The first concern is the one that Helen stated and that just got repeated, that there is some unmeasured fragility or unmeasured need for care in these populations that are not captured in the current risk adjustment variables, and we need to take that into account.

The second is a slightly different

issue, which is, even if there's no greater fragility or unmeasured need for care, because of the limited resources that these individuals have or the neighborhoods that they live in have, the healthcare system has to spend more on them to meet their needs for care, because the resources that we other expect them to have available are not there.

So, the second issue is whether these patients impose more demand on the system for care. And if the system doesn't respond to that, does it fall short in terms of the quality outcomes? So, that's the second issue, which needs to be articulated, frankly, better than I just did it, but also needs to be examined when folks are looking.

The third issue is the fact that we are not completely measuring resources. And the question is, whether the resources that are available in systems that treat a disproportionate number of the disadvantaged population simply have fewer resources and,

therefore, part of the explanation for why they are doing a poorer job is they don't have the resources to do as good a job?

That's not giving them a bye on it, necessarily, but it does recognize that there's a difference between inefficiently using your resources to deliver effective care and not having the resources to deliver effective care.

And so, one of the issues we're trying to deal with here is getting at that, getting evidence or suggestions that that might be the case. So, those are at least the three reasons I see for wanting to have SDS kinds of adjustments and SDS analysis.

And if people aren't responding to those issues or if they aren't responding to questions related to those three goals, then we don't have usable data to assess whether or not the SDS adjustments are needed or appropriate.

DR. AMIN: Those are excellent comments, Jack. And I think, Erin, as you're thinking about feedback back to the Disparities

Committee, the last point that Jack noted might be particularly important for cost and resource use different than quality measures, and conceptually, the issue of disparities might be different in our area of measurement for that reason.

MEMBER WEINTRAUB: It might be different, but it can be important in both. And I think there's a real problem here, because the resources available to less advantaged people is really a latent variable, because it's not the things that you're actually measuring. That's pretty, pretty tricky, but I think Jack, once again, raises an excellent point.

CO-CHAIR DAMBERG: So, this is Cheryl. I think this kind of dovetails back to the comment Kristine made, in terms of thinking about how these measures are used. Because I know CMS has been considering how much they need to do what I call a front-end fix, which is risk adjust, versus a back-end fix, which is to drive more resources to providers who are in your third

bucket, Jack.

And so, I think kind of as these measures get used in different applications, we have to think about sort of the implications. Because risk adjustment is fixing a mismeasurement problem, but it still doesn't address the resource disparity.

CO-CHAIR ASPLIN: Just a quick comment. And maybe this is saying the exact same thing in a slightly different way. I get the sense with all the conversations, especially all the phone calls related to the three measures as we went through this trial period, that there's a belief that there's some data out there that's going to lead to the holy grail of figuring out the differences.

That it's there, we just haven't found the right data elements to put in the risk adjustment model. I just don't think that's where we're going to find it, because we're really, at the end of the day, measuring the outcome variable as cost to the payer, whether

it's Medicare or a private payer.

And the resources that HCMC is expending to have the same level of readmission with its population may be a lot higher than the resources Mayo and St. Mary's are expending in Rochester to have the same level of readmission.

And that's never going to be captured by payments, and maybe that's exactly what everybody else was saying, but I just don't think it's in the risk adjustment model. It might be, it's still open.

MEMBER WEINTRAUB: I think we're all on the same page about that. There are things that you'd like to measure, you know it's out there, but can't.

DR. AMIN: Yes. Okay. Those are very -- I mean, I don't mean to cut the conversation, if there's additional comments, it doesn't sound like there are any in the room, if there's anyone on the phone on this topic? And then, I'd like to ask Ashlie -- go ahead.

DR. BURSTIN: A couple people have

raised this question of the underlying resources that go to different entities, now, that's not measure at all, and I'm not saying necessarily that's a performance measure, but I guess one question as we think about balancing measures or something along those lines, is that potentially a gap to be able to understand the actual resources of the entities being measured as part of this effort? Just a thought to throw out there.

MEMBER GARRETT: If I could just respond quickly? I mean, I think it could be, Helen. And, I mean, it's a really -- actually, there's a simple way to measure it, which is payer mix.

I mean, the percentage of commercial patients you have, that's the way the healthcare system is set up right now, to offset the costs for other populations that don't have -- that have, as HealthPartners said, half the reimbursement or less.

And so, that's actually a simple way

to understand resources and how they're being allocated. And then, you layer on top of that, pay-for-performance programs and all the other ways money is moved around, but that's kind of the basis of the system of inequality we have.

DR. BURSTIN: That would certainly get at the conceptualization of this that's about poverty. I guess, the piece of this we still are grappling with is whether there's truly just something we're not measuring that's different across hospitals that may get at this complexity too.

We constantly get letters from hospitals and others who have raised issues. A recent appeal we just got clearly pointed out in our hospital, a quaternary hospital, this would not be captured. Things like that I think are important to note as well.

MEMBER NEEDLEMAN: On that point, you see, every few years, MedPAC comes out with its report and it shows that hospitals -- there's a Medicare deficit, right? Medicare payments are

less than cost in some hospitals than others, and typically it's the hospitals with large numbers of private payers.

So, you can reinterpret that as saying, hospitals which have the option, have had the option of receiving more funds from private payers are using those to top-up payments beyond those they get from Medicare. It's not a Medicare loss per se, it's a decision to invest resources above the payment level in the Medicare patients.

And other hospitals that don't have that private payer base don't have that flexibility. The finances and the amount of resources per Medicare patient at UCLA-Harbor look very different from those at UCLA-Reagan.

And the issue is, does that affect their ability to deliver quality care? So, some of that was supposed to be equalized by the disproportionate share payments and so forth, but those are the kinds of issues, I think Helen has got it exactly right.

It's not clear that it should be covered in the measurement per se, but we need to have better studies of the actual resources that are being provided to patients versus the standardized payments that are the basis of most of the CMS cost and resource us measures that we've been analyzing.

CO-CHAIR DAMBERG: Srinivas, did you want to make a comment?

MEMBER SRIDHARA: Yes. I mean, I think we're continuing to center around sort of claims-based sort of analyses and so, for the fundamental question I think Helen is asking and Brent also mentioned, if we're talking about actual costs, for health systems or anyone else, that's everything from supply costs to personnel costs.

I mean, there's costs and what we're really talking about is this payment or spending or something else that is unrelated. I think the challenge is, you don't have standardized data assets out there to get across everyone easily.

So, I think it is undoubtably valuable, the question is, how would you do it and do it systematically?

And I think, in as much as NQF has leverage in sort of pushing towards encouraging those sort of democratization of that kind of data, I think that would be meaningful. But without that, I don't know how developers would really get access to that in a way that would be helpful to everyone.

CO-CHAIR DAMBERG: Any other comments on the phone? Okay.

DR. AMIN: So, yes, let's move on. So, I think, just a quick time check, we have half an hour for -- all right. Do you want to do attribution or grouper? You want to do attribution?

MS. O'ROURKE: All right. Let's head to attribution.

DR. AMIN: Okay.

MS. O'ROURKE: So, again, this is something that we just wanted to give you a quick

update on what our Attribution Expert Panel found. Srinivas and Troy were actually on that committee and have joined the Cost Committee, so if there's anything we're missing, feel free to chime in.

Since this has been such a theme throughout this Committee's work, we did want to update you on what this panel found. Again, this slide is some background.

As we continue to see this push to value-based purchasing, it's more imperative than ever that we're able to accurately attribute results to providers, as more and more payment is determined by the results of cost and quality measures.

However, we're finding it's also increasingly challenging to do this, as we're also trying to assess quality on outcome measures, rather than process or structural measures.

So, attribution, we've defined that as the methodology used to assign patients and their

quality outcomes to providers or clinicians. And attribution models can help to identify a patient relationship that can be used to establish accountability for quality and cost.

And then, as we aim to move the system from fee-for-service payment to Alternative Payment Models, it's really highlighted the need to better understand how patient outcomes and costs can be accurately attributed, as we also aim to move to a system built on shared accountability. So, next slide.

So, the purpose of this project was to get multi-stakeholder guidance for the field on approaches to issues of attribution. We aimed to identify key challenges, develop a set of guiding principles, identify the elements of an attribution model and explore their strengths and weaknesses, and then, make recommendations for developing, selecting, and implementing an attribution model. Next slide.

So, here you can see who was on the committee. It was chaired by Ateev Mehrotra and

Carol Raphael. Next slide.

So, we began this project by commissioning an environmental scan from a team of authors out of the University of Michigan and University of Pennsylvania. They found about 163 models that are in use or proposed for use, about 17 percent of those were currently in use. The vast majority used retrospective attribution and, again, the majority, to a single provider, usually a physician.

They found quite a lot of variation in how models are characterized, some of the attributes are on the screen before you. Program stage, the type of provider attributed, the timing, clinical circumstances, the payer or programmatic circumstances, the exclusivity of attribution, the measure used to make attribution as well as the minimum requirement, and the period of time for which a provider is responsible. Next slide.

So, what we found in the commission paper was that largely best practices have not

yet been determined. New models are built off of the previously used approaches and there hasn't been a lot of exploration of the tradeoff of -- the pros and cons of different approaches and that kind of thoughtful consideration of why one model might be better in a given circumstances than another.

One of the real challenges to doing that was a lack of a standard definition for an attribution model, makes it very difficult currently to evaluate across models to determine which one may be better in a given circumstance. Next slide.

So, some key challenges that the committee highlighted. First was that greater standardization among attribution models is needed to allow comparisons between the models and to really allow best practices to emerge.

There's little consistency across models, however, there is evidence that changing attribution roles can significantly alter results and how a provider may perform on a value-based

purchasing program.

And there is currently a lack of transparency on how results are attributed and providers have no way to appeal the results of a model that may wrongly assign responsibility to them. Next slide.

So, to start to address these challenges, the committee came up with a number of products. They first developed guiding principles, they made a series of recommendations, and they created the Attribution Model Selection Guide.

These products allow for greater standardization, transparency, and stakeholder buy-in. And we're hoping to lay the groundwork to develop an evidence base that will allow for evaluation of attribution models in the future. Next slide.

DR. AMIN: Erin, maybe I can jump in here and just talk through some of the principles -- no, you can continue on to the next slide -- just talk through some of the principles. And

then, I'll just walk through the guiding principles and you can walk through the --

MS. O'ROURKE: Yes.

DR. AMIN: -- Selection Guide. And so, some of the guiding principle preamble here was to acknowledge the complexity and the multidimensional challenges with implementing attribution models.

And then, really grounding, and it should be grounded in the National Quality Strategy is, attribution models can play a critical role in advancing these goals. Recognizing attribution models can refer both to the attribution of patients for accountability purposes, but also the attribution of results of a performance measure.

What was pretty critical was to highlight the absence of a gold standard for designing and selecting an attribution model. And, really, the attribution model must understand the goals of each use case.

So, key criteria for selecting an

attribution model are really actionability, accuracy, fairness, and transparency. And I think that's reflective of some of the conversations we've had today around attribution. So, moving on to the next slide.

So, the goals that were -- the guiding principles that were discussed were that attribution models should fairly and accurately assign accountability. That attribution models are an essential part of measure development, implementation, and program and policy design.

And so, this obviously interacts with the conversation we had early today around an attribution model is still a guideline, as it relates to our cost and resource use specifications. That we should consider choices among the available data are fundamental to the design of attribution models.

Attribution models should be regularly reviewed and updated. They should be transparent and consistently applied and they should align with the stated goals and purpose of the program.

And so, again, this relates to Jack's early point around the use of a measure and the measure specification.

And so, the Selection Guide noted that in the current state that there's a tension between the desire for clarity around attribution models fit for purpose and the state of the science related to attribution.

And there's a desire for rules to clarify which attribution models should be used in any given circumstances, but the problem is, there's not enough evidence to support the development of such rules at the current time.

And so, the goal of the Selection Guide was really to aid measure developers, standing committees, such as this, and program implementers on the necessary elements of attribution that should be specified a priori. And it is intended to represent the minimum amount of elements that should be shared with accountable entities. So, Erin, can you just walk us through the Selection Guide?

MS. O'ROURKE: Sure. So, really, the Selection Guide asks a series of questions, and if you look at the committee's final report, there is a little more context here, they go into some of the pros and cons and some of the different things you might consider as you answer each question.

So, Question 1 asks about the context and goal of the program. Again, the committee really stressed that how you're -- what you are trying to achieve is fundamental to your attribution model. And there's a need currently to balance between aspirational goals as well as what's feasible and practical.

So, again, you see in the second column a series of questions that the committee asked program implementers and measure developers to consider as they weigh their choices around attribution.

The next question is, how do the measures relate to the context in which they're being used? Particularly, asking some questions

focusing on the inclusion and exclusion criteria, as well as, does the model attribute enough individuals to draw fair conclusions?

The next overarching heading would be, which units will be affected by the model? Again, some additional questions here on what units are eligible, the degree to which the accountable unit can influence outcomes, again, are there -- do the units have sufficient sample size to meaningfully aggregate measure results, and are there multiple units to which the model could be applied?

And then, finally, how is attribution performed? What data is used, do all parties have access to that data? What are the qualifying events for attribution, do those events accurately assign care to the right accountable unit? What are the details of the algorithm used to assign responsibility? Have you consider multiple methodologies to ensure reliability of the model? And what is happening of the attribution computation?

So, again, the final product, if you will, that the committee came up is a series of recommendations. They are designed to build on the principles and the Attribution Model Selection Guide. They're intended to apply broadly to developing, selecting, and implementing attribution models in the context of public and private sector accountability programs.

The committee recognized the current state of the science, they considered what they can achieve right now, as well as what's the ideal state for the future regarding attribution, and stressed the importance of aspiration and actionable recommendations to drive the field forward.

So, the first recommendation is to use the Attribution Model Selection Guide to evaluate the factors to consider in the choice of an attribution model. The committee stressed that currently there is no gold standard, different approaches may be more appropriate than others in

a certain situation.

The model choice should be dictated by the context in which it will be used and supported by evidence. And developers and program implementers should be transparent about potential tradeoffs between the accountability mechanism, the gap for improvement, the sphere of influence of the accountable entity, as well as the scientific properties of the measure. Next slide.

Attribution models should be tested. Attribution models for quality initiative programs should be subject to some degree of testing for goodness of fit, scientific rigor, and unintended consequences, recognizing the degree may vary on the states of the program. However, given where we are now, models would be greatly improved by rigorous scientific testing and making the results of that testing public.

And when used in a mandatory accountability program, models should be subject to testing that demonstrates adequate sample

size, appropriate outlier exclusion, and/or risk adjustment, to allow fair comparisons, and sufficiently accurate data sources to support the model. Next slide.

Attribution models should be subject to multi-stakeholder review. Given that there is no gold standard right now, the perspective on which approach is best is really influenced by stakeholder interests.

Attribution model selection and implementation in various public and private sector programs should use a multi-stakeholder review to determine what model may best fit their purposes. Next slide.

Attribution models should attribute care to entities who can influence care and outcomes. Right now, models can unfairly assign results to entities who have little control over the patient's outcomes.

And for a model to be fair and meaningful, the accountable unit really must be able to influence the outcomes that it's being

held responsible for, either directly or by collaborating with others.

And as care is increasingly delivered by teams and we're moving to a system of shared accountability, models should reflect whether the accountable entity can influence, rather than directly control.

DR. AMIN: So, Erin, before we move on from this, I just want to link this to our prior conversation around, clearly the tradeoff between an ideal state of what the care delivery system looks like currently and then what our ideal state is, and then, the culture of care delivery that may exist in different communities.

So, that really came up in our total cost of care discussion. But, ultimately, the question of who should be attributed, there should be some ability to influence the outcome. And then, I'll just sort of finish off with the last one.

Which is -- if we can move on to the next slide? Which is, attribution models used in

mandatory public reporting or payment programs should meet minimum criteria. And that goes back to Jack's point earlier in this conversation around, for several of these programs, we're looking at the measure, but we're not looking at the program.

And so, there is an element, particularly for mandatory public reporting or payment programs, that attribution models should be clearly articulated, should be able to -- the accountable entity should be able to meaningfully influence the outcome, adequate sample sizes and exclusions, and several other components.

So, the discussion item that we had here for this conversation was really, as we think about the implications of -- as we think about the outcomes of this Attribution Committee, what are the implications for our work for the Standing Committee?

In particular, the question around, how do we think about attribution models for measure specifications going forward and how do

we think about these concerns in terms of measure endorsement review?

And so, that's really sort of at a very high level what this committee sort of reviewed. I think what might be helpful is we'll sort of share, again, the details of the Standing Committee review and we can discuss it in more detail at a later time. But those are essentially the discussion items that we had for today.

MEMBER BECKER: So, this is Larry, can I ask a question?

CO-CHAIR DAMBERG: Sure. Go ahead, Larry.

MEMBER BECKER: So, I couldn't read through that slide where you had all the committee members and the people I could read, I didn't know all of them. So, pardon me for this. But I wondered whether you had real patients who use the system in various ways as real input and as real partners in putting this work together?

Because it seems to me that

attribution and who has control over what happens to the patient is in large measure what the patient thinks in terms of who they think should be either quarterbacking or whom they are getting their advice, if you will, from.

So, while you might think that it's the primary care doc, as an example, it might be the surgeon or it might be some specialist or -- so, understanding and really working with the patients to understand how they use the system of healthcare, given various sort of types of healthcare issues, I would think would be incredibly important.

And now, you've put up the list and I'm looking at this and virtually everybody on this list has got multiple initials after their names, which tells me that there aren't a lot of patients involved here.

DR. BURSTIN: Yes, Larry, this is Helen. I mean, it's certainly a fair point, we had several consumer and purchaser reps at the table, I guess all of us at some level have to

interact. But I certainly hear your point. The committee did spend a lot of time thinking about how you would actually think about it through the patient lens.

In fact, even looking at some of the newer coding CMS is putting out where it is really assigning even those codes prospectively, they very much said, at the end of the day, ensuring that it's the patient, that's sort of the key thing to always keep an eye on, that they know who is accountable for them. But it's a fair point, certainly more work to do in that space going forward, without a doubt.

CO-CHAIR DAMBERG: Other comments from those on the phone?

MEMBER RAMBUR: I have a question, this is Betty. So, when Vermont was negotiating its all-payer total cost of care model, attribution not surprisingly was a huge issue.

And one of the things that really confounded identifying who's accountable was issues related to nurse practitioners and PAs not

being able to be attributed providers in Medicare, but could in commercial and Medicaid, but also the incident to billing. Did any of that issue -- did that issue come up in this committee's discussion?

MS. O'ROURKE: Yes. I think the committee had quite a few conversations about how you could appropriately include practitioners like NPs or PAs and recognizing the shortcomings that a lot of the current models only assign to physicians. And as we're seeing other practitioners increasingly deliver care, how to appropriately include them in the attribution models.

CO-CHAIR DAMBERG: All right. Any other comments on the phone? Okay.

DR. AMIN: Thank you.

CO-CHAIR DAMBERG: Brent?

DR. AMIN: Oh, sorry.

CO-CHAIR DAMBERG: We have two comments over here.

DR. AMIN: Oh, okay.

CO-CHAIR DAMBERG: Brent?

CO-CHAIR ASPLIN: Yes. So, the framework is very helpful. It's a good conversation. I think, this whole area could be a thousand points of no for us, if we let it.

Because there's always going to be tradeoffs for how granular it is and, I want an episode around not just total knee replacement, but right knee replacement, because that's my sub-specialty.

And we could have a thousand different resource episode measures of things like that. And I just think, we can't let the perfect be the enemy of the good in terms of moving forward in general in this area.

And then, I would just add one other comment around, just theoretically, if you step back from it, a fee-for-service reimbursement system isn't at all designed for accountability around cost.

So, for those providers out there that are really concerned about attribution, should

get themselves on Advanced Alternative Payment Models, because almost all of them deal with this through prospective assignment. Maybe Track One Plus will be an exception.

And there are going to be enough Alternative Payment Models that are specialty-specific or are population-specific for total per capita cost and use prospective assignment that takes care of this.

And I think, just like there's a theoretical limit to risk adjustment, there's a theoretical limit to trying to get attribution right in a fee-for-service environment. And I just, I think if we're trying to get it too perfect, you'll never satisfy everybody.

CO-CHAIR DAMBERG: I'm just going --

MEMBER RAMBUR: This is Betty, if I could just quickly comment? I agree with that. It's more about the base and the trend is often set from the fee-for-service world and the fee-for-service chassis, so that's sort of one of the challenges in moving to real accountability for

cost and outcomes.

CO-CHAIR DAMBERG: I'll make this quick, because I know we're running out of time. I didn't see any mention of unintended consequences. And I know some providers have told me that, like, when the primary care is assigned responsibility, they may offload patients to the cardiologist or wherever. So, I think we need to be cognizant of how attribution rules affect behaviors, because these measures are being used in accountability and payment ways.

The second thing that I would note, and I think in the context of wanting more care coordination, even in this kind of fee-for-service environment, is, I didn't see any mention of multi-provider attribution.

And I think that this comment came up once upon a time in this Medicare Spending Per Beneficiary Measure, because so much of the cost and variations happening in the skilled nursing facility. So, I would, again, raise that.

And then, lastly, related to building the science, I think I agree with that comment, but I think measure developers and their sponsors have to be committed to putting that information out in the public domain.

And I know some measure developers and their sponsors have been reluctant to do that, because sometimes they think they're working in a very policy-sensitive space. So, I would just note that.

DR. AMIN: Okay. I -- Nancy, did you have another comment?

MEMBER GARRETT: So, if I'm understanding this correctly, one of the questions is, whether the attribution algorithm should be specified for measures in the cost and resource portfolio, because we saw an example of one where it was just recommended.

And looking at the guidance from the committee, I really feel like it actually needs to be, because -- a couple of reasons. It's such a key part of how that measure ends up working

and is defined, that to not have guidance on that, but the rest of the measure, you're missing a big part.

But then, there's no standardization around that. And then, also, to have harmonization and standardization of measures across populations, you really need that, to be using the same algorithm.

I am sensitive to some of the previous comments about how it doesn't allow you to take into account local variation, so there is a downside to that. But a lot of quality measures, attribution is built into how you define that measure, and so, I don't think it should be different for the cost and resource measures.

DR. AMIN: So, I think we're going to have to -- we'll take that comment and we'll have to flag it for future Committee discussion. We're going to have to discuss that probably in more detail, but it sounds like there's enough there that we should think about at least moving it into specifications or -- yes, basically, I

think that's the proposal. But we'll have to have a more thorough conversation, given that we're at the top of the hour. Jack, last comment.

MEMBER NEEDLEMAN: Just following up on Nancy's comment. So, I don't have any -- two questions that we need to think about some more. I think Nancy's right, that we -- the attribution is central to a lot of these measures.

So, the question is, two questions, what do we expect the developers to present to us to help us understand the performance, their rationale for the choice of attribution measure and its performance? And today, we simply had somebody say, we used the current state-of-the-art, which is plurality. Is that acceptable?

Which also raises the question, what's our standard for -- can we develop some standards or something beyond our gut feel for what attributions look so questionable that we simply would not accept them? And I don't think we have that answer, we don't have standards right now or

a way to even think about that question.

CO-CHAIR DAMBERG: Troy, I know you have your hand up.

MEMBER FIESINGER: So, this is Troy. I just wanted to chime in to something on the committee. I appreciate everybody's comments, certainly we had proxies for patient representation, we didn't have them -- we just tried very hard to think of ourselves and our patients in that scenario.

We did look at nurse practitioners and physician assistants and Medicare certainly incorporates them. This methodology, we were very aware of the tension between the perfect method and an adequate method, and that's why we looked at minimum standards.

An issue that came up often is, what attribution model is necessary for cost or payment accountability versus what method might be used internally to assign responsibility within the organization?

We talked a lot about unintended

consequences, so that may not have come through in the summary. And, certainly, a lot of discussion about single attribution versus multiple attribution.

The lack of data and research was problematic, just not a lot of this is in studies, a lot of that work was done by people in the room. So, when we try to look at evidence base, there wasn't enough there to draw some conclusions. But I appreciate everyone's input, your comments actually reflect a lot of what we talked about at length in there --

DR. AMIN: Thanks, Troy. Thanks, that's great. A lot of really thoughtful feedback, so thank you all for sticking with us in that marathon of various topics that were on our minds as it relates to the measurement science of cost and resource use. I'll turn it back over to the Chairs for our public and Member comment and then, closing remarks.

CO-CHAIR DAMBERG: So, are there any final comments from the public or the Members?

OPERATOR: Okay. At this time, if you would like to make a comment, please press Star, then the number 1.

CO-CHAIR DAMBERG: Thank you.

OPERATOR: There are no public comments at this time.

CO-CHAIR DAMBERG: All right. Thank you very much.

DR. AMIN: So, I think this is closing remarks from you all, from the Chairs. And then, I think there was some timeline discussions that Irvin's going to walk through

CO-CHAIR DAMBERG: All right. I just want to say, on behalf of Brent, myself, and the NQF staff, really appreciate you guys dedicating not only today, a full day of working through a lot of really complex issues, but also the time you spent preparing for this meeting and continuing to serve on the Committee.

MEMBER WEINTRAUB: Thanks for all the work you guys have done putting this together and making this easy for us.

MS. O'ROURKE: Sure. So, all right. On behalf of the staff, I just want to thank Brent and Cheryl for their leadership today and as we prepared for this meeting. And, again, thank you to all of you for your going above and beyond to come here in person or to hang in with us on a very long web meeting, and bearing with us for the voting technology. So, thank you so much and safe travels home, everyone.

Irvin is just going to quickly run through some of our next steps. I'll steal his thunder and give you the good news that we can cancel next week's post-meeting call, since we got through all of our agenda today. So, taking that out of your next steps.

MR. SINGH: Yes. So, just to sort of echo what Erin was saying, so we had our in-person meeting, which was today. We just canceled the March 22 and March 24 meetings, so we're going to take that off the board.

So, the next active steps in terms of the entire Committee getting together is going to

take place on June 6, 2017, from 2:00 to 4:30 p.m. Eastern Standard Time, and June 8, 2017, from 1:00 to 3:30 p.m., where we're going to discuss the comments that we have received during the public and Member commenting period, which is going to take place between April 20, 2017 to May 19, 2017.

MS. O'ROURKE: So, I think with that, that is all we have today. So, again, thank you so much, everyone. We really appreciate all the time you took to attend and to prepare for the meeting. So, we'll talk to you in June.

(Whereupon, the above-entitled matter went off the record at 4:02 p.m.)

J

Q

S

C E R T I F I C A T E

This is to certify that the foregoing transcript

In the matter of: Cost and Resource Use Standing Committee

Before: NQF

Date: 03-15-17

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