NATIONAL QUALITY FORUM

+ + + + +

COST AND RESOURCE USE PHASE II: CARDIOVASCULAR CONDITION-SPECIFIC STANDING COMMITTEE MEETING

+ + + + +

WEDNESDAY

MARCH 5, 2014

+ + + + +

The 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 Lisa Latts, Co-Chairs, presiding.

PRESENT:

BRENT ASPLIN, MD, MPH (Committee Co-Chair), Fairview Health Services
LISA LATTS, MD, MSPH, MBA, FACP (Committee Co-Chair), WellPoint
ARIEL BAYEWITZ, WellPoint*
LAWRENCE BECKER, Xerox Corporation*
MARY ANN CLARK, MPH, Intralign*
CHERYL DAMBERG, PhD, MPH, RAND Corporation
JENNIFER EAMES HUFF, MPH, Pacific Business Group on Health*
NANCY GARRETT, PhD, Hennepin County Medical Center
ANDREA GELZER, MD, MS, FACP, AmeriHealth Mercy Family of Companies
MATTHEW McHUGH, PhD, JD, MPH, RN, CRNP, FAAN, University of Pennsylvania
JAMES NAESSENS, ScD, MPH, Mayo Clinic
JACK NEEDLEMAN, PhD, UCLA Fielding School of Public Health
EUGENE NELSON, DSc, MPH, Dartmouth Institute For Health Policy and Clinical Practice*
JANIS ORLOWSKI, MD, MACP, Association of American Medical Colleges
CAROLYN PARE, Minnesota Health Action Group
JOHN RATLIFF, MD, FACS, FAANS, American Association of Neurological Surgeons*
ANDREW RYAN, PhD, Weill Cornell Medical College
JOSEPH STEPHANSKY, PhD, Michigan Health & Hospital Association*
THOMAS TSANG, MD, FACP, Merck
LINA WALKER, PhD, AARP - Public Policy Institute
WILLIAM WEINTRAUB, MD, FACC, Christiana Care Health System
HERBERT WONG, PhD, Agency for Healthcare Research and Quality
DOLORES YANAGIHARA, MPH, Integrated Healthcare Association
NQF STAFF:

HELEN BURSTIN, MD, MPH, Senior Vice President, Performance Measurement
TAROON AMIN, MA, MPH, Senior Director, Performance Measurement
ANN PHILLIPS, Project Analyst
ASHLIE WILBON, RN, MPH Managing Director, Performance Measurement
EVAN WILLIAMSON, MPH, MS, Project Manager, Performance Measurement

ALSO PRESENT:
BENJAMIN N. HAMLIN, MPH, NCQA
BOB REHM, MBA, NCQA
ROBERT SAUNDERS, PhD, NCQA

* Present by teleconference

T-A-B-L-E O-F C-O-N-T-E-N-T-S

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

(9:02 a.m.)

MR. WILLIAMSON: Good morning, everyone, and welcome to Day 2 of the Cost and Resource Use Standing Committee meeting. I want to thank everybody for joining for us today and thank everybody for their participation yesterday. I think we had a productive day with some strategic discussions as well as measure evaluation.

At this time we'll turn it over to our co-chairs, Brent and Lisa, and we'll take care of a few disclosures this morning and then do a quick recap of yesterday and then dive right in.

DR. ASPLIN: Very good. Thank you, Evan. Good morning, everyone. I'd like to welcome Tom Tsang, good to see you. And I wonder if you could introduce yourself to the committee, and if you have any conflicts disclose those for us.

DR. TSANG: Yes, this is Tom

Tsang, executive director at the Merck Medical Information and Innovations Group, but no disclosures.

DR. ASPLIN: Great. And everyone else was here yesterday, I believe. Are there any committee members that are attending by phone that did not have an opportunity yesterday to declare any potential conflicts or disclosures?

MR. NELSON: Gene Nelson is on the phone, and I was not able to attend yesterday.

DR. ASPLIN: Welcome Gene. Do you have any disclosures for the committee?

MR. NELSON: Let's see, yes. I'm at Dartmouth, at the Dartmouth Institute, and we do a great deal of research on costs and the value of care. I am a founder of a quality measurement company which sometimes includes value assessments and reporting that's called Quality Data Management. And I think those are the major potential conflicts.

DR. ASPLIN: Thank you Gene, I

appreciate that. Has that company done any evaluation on any of the NCQA measures or specifically the measure that we will be discussing this morning?

MR. NELSON: No.

DR. ASPLIN: Great. Any questions from committee members for Gene or Tom? Any updates or announcements this morning before we get started?

From a housekeeping standpoint, Evan, one question I had was has the dates for the next in-person, have those been communicated?

MR. WILLIAMSON: Yes, we will be discussing the next steps and committee timeline at the end of the meeting today, but all the dates for Phase II and Phase III have been set. So we'll go over those and make sure everybody's aware of the responsibilities for the committee.

DR. ASPLIN: Very good. Sounds good. With that we are going to move forward

this morning with a quick overview of the day. I think we've completed our recap. We have the NCQA measure in front of us, have time for public and member comment following our consideration of the third measure, and then this afternoon we'll have kind of a continuation of the dialogue we began yesterday morning around the future direction for cost measurement and what Phase III for the project will look like.

Again have opportunity for member comment and public comment and discuss the timeline before adjourning. So that's the outline for the day, and let's get started.

So with that we have Measure 1558, the Relative Resource Use for People with Cardiovascular Conditions from NCQA. This is an endorsed NQF measure that is up for reconsideration by the committee.

On the phone with us today from NCQA we have Ben Hamlin. Ben, welcome.

MR. HAMLIN: Thank you.

DR. ASPLIN: Great. Are there other representatives that would like to take a seat at the table here? Introduce yourselves.

MR. WILLIAMSON: Ben, this is Evan. Has the phone line issue been resolved?

MR. HAMLIN: Yes, I think it was just some feedback from one of the other members. It's fine now.

MR. WILLIAMSON: Okay, great. Yes, just let me know if at any point the line goes out or you can't hear us.

MR. HAMLIN: Okay, thank you.

DR. ASPLIN: Evan, could you look at the list of committee members that are online for everyone so we know who all is online?

MR. WILLIAMSON: Absolutely. Right now, logged into the webinar we have Ariel Bayewitz, Gene Nelson, Joe Stephansky, John Ratliff, Larry Becker, and Mary Ann Clark.

So Gene, I know you weren't on yesterday, so we'll go over a bit of a process step here. There is a chat feature associated with the webinar that we are using kind of as a virtual placard raising if you would like to speak. So just send the leaders a message at any point you want to make a comment and we'll let you know that you're in the queue. So that's how we'll handle the remote participation.

MR. NELSON: Sure.

MR. WILLIAMSON: We also have voting set up through the webinar, and I'll discuss that before the first vote just to make sure we go over that again. I know we have some new members in the room here as well as on the webinar. So we'll make sure that everybody's clear as to what the voting process is. So thanks for that. We'll turn it back over to Brent.

DR. ASPLIN: Great. And so for this measure, first we'll have an opportunity,

Ben, for you to introduce yourself and colleagues here that are with us in the room. We'd ask for a brief introduction, an overview of the measure, and if we could keep that introduction at a high level and less than five minutes that would be great.

We then have two lead discussants from the committee, Andy Ryan and John Ratliff, who will give their assessment of the comments that the committee submitted online prior to the meeting, highlighting areas of both agreement and potential disagreement.

For Andy and John, we're going to do that by category. So we'll start with importance to measure, then move through scientific acceptability, feasibility and usability. And then of course Bill will again represent us from the TEP.

And with just the sheer discipline that we demonstrated yesterday of keeping our comments to the section that we're voting on we'll move through the rest of the sections

this morning.

Good. Very good. So Ben, my understanding is that you'll be taking the lead, so if you could take a moment to introduce yourself and then we'll have your colleagues in the room do the same.

MR. HAMLIN: Sure. I am Ben Hamlin and I am the director of Performance Measurement at NCQA and I'm also the project director for the Relative Resource Use Measure Domain and Efficiency Measures at NCQA.

MR. REHM: Hi, my name is Bob Rehm. I'm Assistant Vice President for Performance Measurement at NCQA.

MR. SAUNDERS: I'm Robert Saunders. I'm Assistant Vice President for Research and Analysis at NCQA as well.

DR. ASPLIN: Great. Welcome, and I'll turn it over to Ben. Great, thank you.

MR. HAMLIN: Okay. So our Relative Resource Use for People with Cardiovascular Conditions measures how

intensively health plans use resources in managing their members with a specific list of cardiovascular conditions identified through claims.

This measure uses standardized prices that are published by NCQA, actively creating a process by which health plans can compare their total annual resource use to their own peers in a meaningful manner.

NCQA receives aggregate data submitted by plans which is verified by NCQA certified auditors, and then NCQA uses all plan submissions to calculate national and regional benchmarks for all plans in addition to individual specific plan benchmarks for each of the service categories that are displayed for the RRU measure.

This enables health plans to understand how their own resource use for their members with chronic disease compares both to their peers and also to others across the U.S.

NCQA presents the observed resource use data along the calculated benchmarks for each service category for each plan, and that again allows them to compare their observed resource use to the calculated benchmarks.

The national and regional results for each plan are presented alongside a HEDIS quality composite in order to create a value equation that the plan provides to their members with chronic disease.

We found these measures are of increasing interest to consumers and employers and government programs, helping them identify the best value and the high quality care that's delivered most efficiently and cost effectively.

So I'd be happy to answer any questions that the committee may have.

DR. ASPLIN: Thank you, Ben. Appreciate that. I think we will have plenty of questions as we move through, so unless

there are any quick, high level issues I'd like to hear probably first from Andy.

And Andy, if you could focus on the Importance to Measure section that would be great.

MR. RYAN: Sure. So the committee -- I think there was wide agreement that this is a high priority area. It's important to measure with respect to opportunity for --

DR. ASPLIN: Andy, could you move your mic just a little closer, please? I'm sorry. Thank you.

MR. RYAN: Sure. With respect to opportunity for improvement, I think there was general agreement that the developers' explanation was okay. There was some question about there not being evidence about variation in performance across plans.

There was a couple points made that there wasn't data from the point at which the measure had originally been endorsed a couple years ago. All the data shown were

quite old.

And then there were other questions about the fact that the measure is intended to be at the plan level and having some, this was throughout the comments, but raising some question as to whether, you know, assessment at the provider level would provide greater potential for improvement. That had been mentioned.

And then also there wasn't information on disparities shown from the developers. Those were points that were made with respect to importance, but these -- my read wasn't that these were huge problems, just kind of requests for more information from the developer.

DR. ASPLIN: Very good. Thank you.

John, do you have comments around the Importance to Measure, Measure Intent, the first category?

DR. RATLIFF: I think that

summation was very good. At least from our comments there was uniform support for the priority of the measure.

With regards to the opportunity for improvements, one of the commenters did note that using unit of analysis in the health plan might be suboptimal and using that information with regards to assessing providers would be kind of getting one step away from the level of measurement that you desire with regards to the intent.

And both with regards to the intent and other aspects, there were multiple commenters that brought up data and how this plan has been used over the two years that it's been endorsed, what's been learned from using the measure or what kind of improvements have been engendered because of the measure. That was brought up in multiple sections of the commentary from the standing committee.

DR. ASPLIN: Thank you, John.

Bill, do you have an overview from

the TEP's perspective?

DR. WEINTRAUB: I do. And once again I think the best thing to do would be to display the document that shows the TEP summary and that response. And I probably went through it a little too fast yesterday so I'm going to slow down just a little bit.

The very last portion of it really gets to validation and we come back to the TEP at that time. In each, where there was a question, there was also a developer response which I can either summarize or it might be better to have the developer comment as you'll see it before you. Page 4, middle of Page 4, there we go. Okay.

So the first one, based on stated intent to what extent is the measure appropriate? And there the TEP quite simply felt that the measure was clinically appropriate.

MR. HAMLIN: I'm sorry, could you repeat that? Your words got garbled.

DR. WEINTRAUB: Okay. Sorry, I'll try again. Can you hear me okay now?

MR. HAMLIN: Yes, just the last sentence.

DR. WEINTRAUB: Okay. The TEP agreed that the measure population was clinically appropriate.

MR. HAMLIN: Okay, thank you.

DR. WEINTRAUB: Okay, next. Next, to what extent will the definitions to identify the population clinically consistent with the intent? The TEP was concerned that not all applicable diagnosis codes identifying the population intended were included. And you can see the developer response.

Do you want to comment or shall I summarize? Want me to summarize? Oh, go ahead.

MR. REHM: Ben, do you want to, do you have any summary on the diagnosis question? I think we supplied the value sets.

DR. ASPLIN: Ben, this is Brent.

If we could just have Bill finish the whole TEP summary, and then I'd like to vote on the importance and then we're going to walk through.

Like most measures, I think, we're going to spend most of our time in the case, probably, with the reliability and section of scientific acceptability along with validity, and so we can have a lot more back and forth in that section.

DR. WEINTRAUB: Would you rather then that I summarize the TEP response just to move this along?

DR. ASPLIN: Yes, why don't you finish the TEP responsibility and then we can move forward.

DR. WEINTRAUB: I mean the developer response to the TEP points, so for the second one, the developer response adoption of ICD-10 codes and updates would address this and the TEP agreed.

Okay. The third one, to what

extent does the measure accurately describe the evidence, and the TEP agreed with the developer's logic and grouping claims. They felt that the exclusion of cardiovascular patients with HIV or cancer was of some concern.

The developer's response that the exclusion was based on disproportionate resource use and a plan with a larger number of cancer patients will have results capped out, and overall that the TEP was satisfied with the response.

Okay, fourth question. Given the condition being measured, describe the alignment of the length of episode. The TEP felt that that was appropriate. Fifth question. Describe the clinical relevancy of exclusions. TEP was satisfied with that. Do the exclusions represent a large number of patients? The TEP requested more detail, and the developers said they will present distribution data to the committee. So I

trust that we will see that today.

To what extent is the rationale for clinical exclusions adequately described? There was some concern that the measure of patients excluded from the measure that we're still using resources and a plan that refuses to pay for those resources could appear to be performing better but it was beyond the scope of the evaluation.

The developer responded that this issue is being handled through NCQA accreditation standards, and so I'll just leave it at that.

To what extent are relevant conditions represented in the codes? The TEP was concerned that not all applicable codes were included, and the developer's response was to reevaluate on an ongoing basis.

The next one, to what extent are covariates included? And that really gets to validation. Why don't we come back to the TEP at the time? That's much longer.

DR. ASPLIN: Thank you very much, Bill. I appreciate that. And we're just going to move forward with the first four votes in our evaluation of the measure, and then likely get to the heart of our discussion here.

So the first question, importance to measure and report. You see the question in front of you. And this section relative to this measure is open for discussion. Dolores?

MS. YANAGIHARA: I'm not exactly sure when to bring this up, but I just have a question for NCQA. You know, many people commented on the value of bringing this measure together with the quality measures, but because of the change, the recent change in the LDL guidelines and the recommendation by NCQA that's out for public comment to remove the LDL screening and control measures, I'm just wondering what your thoughts are and what the quality measure would be that would be paired with the resource use measure.

MR. HAMLIN: So the quality composite is comprised of all current HEDIS cardiovascular measures that are eligible for public reporting. And so any modifications to the LDL measure or other quality measures would be reflected in this quality composite should they be approved by our committee on Performance Measurement.

So it should be up to date as of the next publication of the HEDIS measures if those changes are approved.

DR. ASPLIN: Ariel, is your question relative to importance to measure and report?

MR. BAYEWITZ: Yes.

DR. ASPLIN: Great.

MR. BAYEWITZ: So my question, so I don't debate that the relative resource use for people with cardiovascular conditions is high priority. What I just wonder about is how important is it to evaluate this at a plan level? And so when I see a lot of evaluating

plans I think about, you know, so quality makes sense because the intent there is that you think the plan themselves can do something about it, you know, be it in their selection of their network or in terms of their actual engagement directly with the member or they're setting up certain value based purchasing programs with providers to manage that quality.

When I see that resource use though, I was just thinking so what do we expect the plans to do about it? Assuming that we say it's reliable and valid, do we expect them to change medical policy around certain resources so that we can limit resources for people with these conditions? I mean is that the intent of this?

Are we saying that we think that they should help manage, you know, in terms of the selection of their providers they should have a more narrow network or kick some providers out? And if it's the latter, if we

believe that it's the providers that control more of the resource variability, then I would think that we would really need the ability to drill down one layer below this and to be able to evaluate providers.

It doesn't have to be physicians, but even relatively mid-size organizations. And it just seems from reading through the documents that that was not in the scope here, that wasn't really possible.

So I guess my question again, it's not that relative resource use for cardiovascular conditions is not meaningful, but I do question how meaningful it is for a purchaser or a consumer to see this.

And just even, you know, and one step beyond that for a purchaser, when a purchaser is looking at plan quality that's one piece. When they're thinking about the next piece, I think from a purchaser standpoint they're really interested in cost, right? What's it going to cost me? Again, I

don't know if they're thinking medical policy or provider network there around the resource use side.

MR. HAMLIN: So I mean, I think the reason these measures have come into play, I mean these measures have been in development for some time, was the fact that up to the point of where these measures were available the only thing the purchaser had was, you know, the cost of the benefit they were purchasing. There really was no additional information that they had about the value the plan was offering for that cost.

You know, the reason that we use standardized prices in these measures is because there's so much market variation and there's so much variation to cross contracts within each plan, in order to try and create a plan-to-plan comparison metric you do need to address a little bit of that without overdoing it.

And this was the approach that we

have found provides a way for plans to compare themselves to their peers both on resource use and on quality through the use of these metrics.

We don't, through this approach, I'd hate to say judge, but we don't judge plans who have higher resource use in certain categories, necessarily. You know, we just basically present their data compared to their peers in as detailed a fashion as conceivably possible without being able to dive down into some contractually prohibitive data that creates issues for the plans.

And again, you know, given that this is a national comparison strategy, we wanted to be as relevant as possible given the limitations of the measurement approach based on the data available.

So we do not expect that plans will be limiting resources based on their results from this measure. What we do expect plans to do is compare their resource use at

multiple levels to their peers and then drill back down into their own data looking for opportunities and cost opportunities to improve, you know, based on the value that they're seeing from these measure results.

These measure results are fairly high level, I admit that, even though there is a fair amount of detail in them. But again, you know, this is a national plan-to-plan comparison strategy that allows states and employers to sort of understand how plans perform against each other.

The plans themselves will have to do the really heavy lift in drilling down into their data to look for those opportunities specifically.

DR. ASPLIN: Ariel, this is Brent. I would just make a comment from the Twin Cities market that the plans there have taken architecture of this and similar measures and gone to the next level and are reflecting back to delivery systems, both the relative

resource index and also the pricing.

So that's happening in that market. It's not the exact measure that's in front of the committee today, but it's the same architecture of that committee and that has been very helpful feedback as a delivery system leader in that market.

So I have Jack, then Andrea, Carolyn and Dolores.

DR. NAESSENS: Brent, that comment spoke directly to the question I wanted to ask the developers which is clearly NCQA is geared to comparing and providing information at the plan level. That's the rationale for the organization and its contribution in this space, or at least one of its contributions in this space.

But is the coding data, is the methodology available to the plans to do comparable analysis down to the group level, the market level, the physician level so that they can have the opportunity to do the kind

of drill-down we're talking about to make it usable within the plan? Because you've discussed that as one of the goals for the plans in having the data available.

And to what extent do you know whether they have been doing that?

MR. HAMLIN: So all of the methodology that we use for the measure calculation is available to the plans. We provide them back as we mentioned, individual benchmarks for each plan is calculated from the data received so we try and provide them as much information as possible.

I mean we do expect that the plans would, because they have to map all of their resources to the standard pricing they can actually, effectively, use that same methodology and plug in actual cost to do their opportunity-cost calculations, and I would expect plans to do that.

And we do hear stories, as I think Brent just gave you, about different systems

that are sort of using this framework to, you know, provide additional information. We try not to be too prescriptive in how they should be going about doing that.

We do try and offer some suggestions and some stories that we hear back and forth, and again we make our methodology transparent and we publish, you know, again all of our standard pricing tables and the measure of methodology and all of that to try and provide the systems as much information as they possibly need.

There are several demonstration projects where this has been applied to the provider group level, you know, we do hear some success. Because of the complexity of the calculation it does require sort of an organizational level, like NCQA approach, but I do think that it is valuable if you drill down.

And as we move forward we will continue to investigate, you know, taking it

down to the next level if it is, in fact, possible.

DR. ASPLIN: In fact, it might be less valuable if we force them to take it down to the provider system level, because doing so we'd have to use the standardized pricing. And one of the most powerful aspects of having the conversation within the market is that once they go below the level of the standardized pricing they can use actual pricing without disclosing what those prices are, and that's actually very powerful in those conversations.

Let's see. Andrea?

DR. GELZER: Thank you. We do Medicaid-managed care and we're in about 14 states, and we're also, if the dual demo projects ever start we'll be doing those as well. So this measure, not so important in the Medicaid population but hugely important in the dual-eligible space.

And when I first, you know, was

reviewing it, and we don't use it a lot in Medicaid, but when I was first reviewing it for this committee I was thinking, well, you're not going to see variation in a market. You're going to see it market to market. So I was trying to determine, well, is this really even valuable?

But I think as, you know, we're growing rapidly and a national company now. And I think it's valuable to go into a market when you have disparity from market to market, this is, if I have this information, it's valuable for me to go in then and have the discussions with the systems and the provider groups in the higher markets. So I do see value to this measure.

DR. ASPLIN: Thank you. Carolyn?

MS. PARE: I think it's particularly important to note that some of the discussions we had yesterday around who is your audience and who do these measures serve right now, if you look at -- and I'll just

speak to the employers as purchasers.

There's not a whole lot of transparency quality information for them, and employers are always challenged with the fact that they buy based on price and access and never quality. And so NCQA's attempt to somehow convey quality at whatever level they can back to the purchasers allows the purchasers to buy on some indication of quality.

Now that's a point in time sort of thing. If at some point we are directly dealing with provider information and it's transparent and clear for people, Brent talks about the fact, and we talked about this yesterday too. All the contracts are proprietary and so plans and providers can't disclose this information.

Until we have full disclosure of the price and quality, we're going to have to use some kind of proxy for these and that's why this particular measure is so very

important because it allows right now the purchasers to see quality at some level.

DR. ASPLIN: Thank you. John, did you have a question?

DR. RATLIFF: Just a couple quick questions for the developer following up on the comments. The measure's going to be used or is eligible for use with Medicare Advantage plans in their Five-Star system for ratings?

MR. REHM: This is Bob.

MR. HAMLIN: Am I next?

MR. REHM: Ben, if you want to, go ahead. But the RRU measure currently is not in the Stars program.

(Off the record comments.)

DR. RATLIFF: Sorry, I'm off mute. Sorry about that. I think the crackling was me. So this isn't being used in Medicare Advantage?

MR. REHM: We evaluate Medicare Advantage plans in the RRU. We evaluate commercial plans, Medicaid plans and Medicare

Advantage. But CMS decides what measures go into the Stars rating and that is not one of them yet.

DR. ASPLIN: Thank you. Ariel had a follow-up question regarding the minimum number of members per condition for the measure to be meaningful, and specific with the Twin Cities I don't recall whether the number of members with each condition was discussed.

We had enough attributed members that it wouldn't have been a problem with this one. Maybe some of the other diagnoses, I'm not sure we met that threshold. But perhaps the developer, we'd have to ask the plan, so I'm not really sure how you would respond to that.

MR. HAMLIN: So our minimum number for this measure is 250 members. And for this measure we had fewer problems with small sample sizes than we do with some of the other RRU measures, certainly.

MR. BAYEWITZ: What was the 2,000 measure reference that I saw?

MR. HAMLIN: I'm sorry, I don't know what you're referring to.

MR. BAYEWITZ: In one of the documents I thought it talked about -- I'll take a look. I thought there was a mention of a 2,000-member requirement, but I could just be misremembering.

So it's basically saying you need 250 members per condition, and then based on the prevalence of that condition you'd back into what would be the necessary size of the organization for you to evaluate them on this particular measure. So yes?

MR. HAMLIN: Right. So we validated the risk adjustment for this specific measurement approach, you know, to require to our level of comfort that the organization have at least 250 members in the eligible population in order to report the measure, and that holds.

DR. ASPLIN: Thank you. Dolores, do you have a follow-up?

MS. YANAGIHARA: Yes, I just wanted to share that we have actually tested this very measure at the physician organization level in California and it does work. Not all of the physician organizations got results because they didn't have a large enough population, but a majority of them did.

We ultimately didn't end up using the measure in our program because we had other utilization and cost measures that we could use, but it definitely was of interest to the plans to try to get to the next level and, you know, our committees felt like it had valuable information.

Like I said, we just had other measures that we could use that gave more information, but it does work at the physician organization level.

DR. ASPLIN: Thank you. I would like to call the question then on importance

to measure report 1(a), high priority. The options are in front of you. High, moderate, low, and insufficient evidence. And those of you online, we'll set up the online voting, and Evan, could you let us know when you're ready?

MR. WILLIAMSON: So as a refresher you have a vote-snap device. Please direct to the laptop. It's a line-of-sight feature. The numbers correspond to the responses. For the webinar online you will see four options appear when I change the slide. Please select the appropriate response.

I will now vote on high priority. This is subcriteria 1(a) for importance to measure and report. You have four options. You may begin voting now.

I believe we're still waiting for one response in the room. If everybody could please point their device again. One of these days we'll get this right. Yes, there we go.

We have all the votes. And it

looks like we have an issue with our screen in the room. So I believe we have 20 high and two moderate.

DR. ASPLIN: For those of you online we just have a little issue with the screen resolution here. It's not a question of whether we passed that.

All right, we're going to move to 1(b), opportunity for improvement. Demonstration of resource use for cost problems and opportunity for improvement. It's the data demonstrating variation in delivery of care across providers or population groups. So open for comment.

Cheryl?

MS. DAMBERG: Yes, I was struggling a bit in the documentation provided because of the normalization that's done each year. You noted that you can't actually trend the information. So I was trying to figure out how do you gauge whether a plan has improved over time?

MR. HAMLIN: So the issues with these measures are that we utilize all plan submissions to calculate our benchmarks every year. Also the standardized prices are updated every year, and so those benchmarks are basically dependent upon the plans that submit and, you know, the prices.

And so in order for us to actually track a single plan's improvement over time we would have to hold a number of things artificially constant in order to do that.

So again, this is a relative snapshot of a plan's comparison in that year to its peers and, you know, there are some limitations to sort of tracking improvement specifically at NCQA's level, but that doesn't again prohibit a plan from going down to the next level on their own and tracking their own improvement in the services categories.

You know, there are some values, you know, in the frequency of services category that perhaps plans can watch numbers

change, but again whether they assess that as a positive or a negative factor depending upon how the rest of the designs that have links together, that really is up to them to determine.

MS. DAMBERG: So I'm just curious. Has NCQA, I realize there are a lot of moving pieces so it's hard to compare year-to-year, but do you track whether plans actually do shift positions? So, you know, maybe they're above 1 for two years running and then they shift below? I'm just kind of curious.

MR. HAMLIN: We do look at the quadrant shifts for plans, you know, in annual analysis and we sort of do a plan stability analysis to, you know, to see. Around the mean though, you know, it's difficult, because shifting around the mean can be not really all that relative. It's more of the larger shifts.

DR. ASPLIN: Thank you. I have Taroon, then Bill, then Lisa.

MR. AMIN: I just wanted to point out based on Andy's comments yesterday about the differences between maintenance measures and new measures submitted. So opportunity for improvement would be one of the areas where we would expect to see data on the opportunity for improvement with the measure as specified.

And also I'll just point out for reference that additionally the criteria for around usability and use we would expect to have some information about the measure in use.

DR. ASPLIN: Thank you. Bill?

DR. WEINTRAUB: So this is pretty tricky. Clearly we're spending too much money in cardiovascular care. There are places that we're doing things that we shouldn't be doing. We also have tremendous healthcare disparities, so some places we're spending too little on healthcare.

And so what are our goals here?

Our goal is to spend less? Is the goal to reduce variation between plans? Is our goal to decrease disparities in care? And if we reduce what we spend overall, how do we do that without sacrificing quality, and yet we had our discussion yesterday showing there's not a very good relationship between what we spend and quality. So I think there's opportunity for improvement, but getting at that, you know, and you have a good metric for success, I don't think is a small task.

DR. ASPLIN: Thank you. Lisa?

DR. LATTS: So my comments are similar to Bill's. My issue with this measure is always that I don't know what opportunity for improvement means. Other than being clustered around 1, I don't know if as a health plan I want to be high or low.

As an employer maybe you say, well, I want you to be low on this measure, but as a patient I want you to be spending all your resources on me if I need them. So it's

really far less about cost than appropriateness.

So that's just, you know, I think this is important. I think it's a piece of the puzzle but, and again this goes back to my comment from yesterday. We talked about cost and quality. The third leg of the stool is appropriateness, and we just don't know.

DR. ASPLIN: Thank you. Seeing no other requests -- oh, are you good? All right, I'd like to move ahead with voting on Criterion 1(b), opportunity for improvement. Evan, go ahead when you're ready.

MR. WILLIAMSON: We will now vote on Criteria 1(b). I'd like to point out we now have seven voting members on the web, so the numbers will now be out of 23. You can begin the voting now. And we have all the votes. Okay, so we have seven high, 14 moderate, two low, and zero insufficient.

DR. ASPLIN: Thank you.

Next we have Criterion 1(c),

measure intent. Intent of the resource use measure and measure construct are clearly described. Any comments or questions? Seeing none, let's go ahead with voting.

MR. WILLIAMSON: We will now vote on Subcriteria 1(c), measure intent. You have four options. You will begin voting now. And we have all the votes. And we have 17 high and six moderate.

DR. ASPLIN: And overall importance to measure and report considering all three of the votes we just took, Evan, go ahead.

MR. WILLIAMSON: We will now vote on overall importance to measure and report. You have four options, high, moderate, low, or insufficient, and you will begin voting now. And we have all the votes.

DR. ASPLIN: We have some mystery numbers for those of you on the web that you're not seeing, but the bottom line is it's a strong majority that have voted either high

or moderate. So we are going to move on to the next category while we work through our technical details here.

MR. WILLIAMSON: We'll pull the numbers during the break and update the record.

DR. ASPLIN: I think it's 21, or 22, excuse me, between high and moderate and then one rated it as low. So it passes on importance to measure and report. We'll get those subcategories for you.

Next we're going to move forward with scientific acceptability considering both, the two votes, one on reliability, one on validity. And we'll again turn to our lead discussants. Andy, go ahead.

MR. RYAN: Okay. I would like to verify that the documents submitted by NCQA, the only one that has bearing on this question is called SA Reliability, underscore -

MALE PARTICIPANT: Can you speak up please?

MR. RYAN: -- from 2005. Is that the only document that NCQA submitted with respect to the reliability and validity of the measure, the SA, underscore, reliabilities and, underscore, validity from 2005?

DR. ASPLIN: That was a supplemental -- go ahead, Ben.

MR. HAMLIN: Yes, that was a supplement to the measure testing form information that was included as part of that, the 23 or 24-page document that was submitted as part of the measure.

MR. RYAN: Okay, thanks.

All right, so to just give an overview. I think the committee with respect to specifications raised some questions about risk adjustment and how this RRU-HCC risk model differ from the CMS-HCC model in terms of the comorbidities included.

There were, I think, it may be one or two points raised about the specifications with respect to the clinical diagnoses that

identified cardiovascular disease.

I think the largest issues were about reliability and validity testing. I think the kind of methods that the committee is used to seeing that's documented in, say, Algorithm 2 with a signal-to-noise ratio or split-half correlation, the committee wasn't satisfied with what was presented that it showed reliability.

All I could see were standard errors that were shown. And so I think there's just a lack of information about what they did to test for reliability. I think the same thing with validity that there weren't comparisons with other measures to show that, or some, you know, external validation that show -- resource use.

I think, you know, people thought it makes sense. It has some face validity, but in the extent of testing, I think, was lacking. So that's how I would just quickly summarize the comments of the committee.

DR. ASPLIN: Thank you. John, do you have further comments for the committee from your perspective?

DR. RATLIFF: I think that summary of the committee's comments is extremely good. I mean multiple different contributors borrow from the fact that there was insufficient data presented to develop an opinion as to the reliability or validity of the measure. And that was something that echoed through multiple different commenters.

DR. ASPLIN: Very good. So we're going to open up beginning with Bill. Do you have comments, please?

DR. WEINTRAUB: From the TEP, if you could pull up Page 5. That would help. Okay, so our comments were very similar to Andrew's, remarkably.

So there was concerns about both reliability and validity. There was no R-squared that we could find in the materials that we were given. There was concern about

the risk adjustment models applied but not validated.

The response beginning on Page 5 and going on to Page 6 is long and I don't think I should try and summarize the developer's response here. They should.

DR. ASPLIN: Very good. So perhaps we could start with a comment and response from the developers just in this whole space about the lack of specific testing and the concerns raised by the TEP and the committee, and then we'll open it up for dialogue.

MR. HAMLIN: Okay. So the original testing for validity of the measure was primarily, principally, outlined in the document of 2005 and that's when the measure was first, you know, tested for appropriateness in this space.

The 2008 document that was provided in the testing form, the information there was when we did the validation of the

HCC model applicability and appropriateness for the RRU. And that information, I believe, was provided in the NERC measure testing.

These measures are tested annually, so again our continued reliability testing principally is around the identification of outliers over the almost 1,000 plans that submit these measures to NCQA.

And, you know, we look for outliers. We look for errors in the submissions through our audit process. And we, you know, compare the results using fairly extensive correlations, looking at the different service categories to try and identify any areas where the measures, you know, don't conform to what we're seeing.

Unfortunately the limitations of the amount of information we can provide through the actual measure testing form, I think, was scattered, and hence the number of different, rather extensive attachments.

So again we can't do an individual R-squared, our response, I guess, on the data submitted to NCQA annually because we only receive aggregate data from the plans that is verified by the auditors. However, we don't get patient level data submitted by the plans. So, you know, it's because the individual members are already included in the cohorts.

So again, we did test the appropriateness of the HCC model at the patient level using many simulations of patient level data and that was, I believe it was the 2008 document that was submitted. And again, we utilized the HCC approach.

DR. ASPLIN: Cheryl?

MR. HAMLIN: It was developed by CMS

DR. ASPLIN: Sorry.

MR. HAMLIN: -- the supplement that was looking at the validation of the HCC model to resource use measures.

DR. ASPLIN: Cheryl, before you

go, Robert?

MR. SAUNDERS: Sorry. Thanks. The model, the testing that he's describing is built off of the Optum data warehouse. And so the underlying testing has information about individual, has individual level performance information, and so all the risk adjustment testing has been done at that level.

So we will look through our materials to see if that's available within there to report out, but we've essentially taken a model built off of that testing to then apply across all the health plans that submit to us.

DR. ASPLIN: Thank you. I have Cheryl and then Lisa.

MS. DAMBERG: I was wondering if the measure developer could comment on one of the exclusion categories that you note. And this was on Page 11 of your documentation. It says the claim on the service was rejected because it was missing information or was

invalid for some other reason.

And I wasn't sure that I saw in your documentation what proportion of claims fell into that category, because I could imagine that could be quite large, and I have some concern about setting aside those types of claims.

MR. HAMLIN: So there is two parts to that answer, I think. The first is the HEDIS health plan accreditation standards cover the processing of claims and the approval of claims, and I'm not really the person qualified to speak about the, in any kind of detail about those.

However, all HEDIS reporting is based only on claims that are in fact paid by the plan, so that is the limitation that we can work with as far as the submission. And I would agree that perhaps there's some additional information to be gained from those others, but unfortunately those are not accessible to us.

MS. DAMBERG: So you don't actually know the proportion of claims that fall into that category?

MR. HAMLIN: Actually we don't have access to that information. It's only available to the plan, because each plan's different.

MS. DAMBERG: But do you know if that varies across plans?

MR. HAMLIN: Specifically no. I would expect over, nationally I would expect they're similar enough, but I don't have the data to make any kind of assertion in that regard.

DR. ASPLIN: Thank you. Lisa and then Nancy?

DR. LATTS: So my question is actually for NQF. Since this is a recertification, do we, you know, so what's sort of the expectation of the developers in terms of testing?

Are they expected to do sort of

again de novo or can the committee accept what was presented the first time around when it was originally approved, or what's the sort of expectation of the developer?

DR. BURSTIN: This has been an issue we've talked about a lot. At this point we have not required additional testing. We'd love to see it, particularly when a measure's out in use, but it's not something we can really require. Especially when we recognize, you know, three years sounds like a long time to some of us, but in the world of actually putting a measure into place, finding out about its use, it's a lot quicker than we think.

MR. AMIN: Well, I would just clarify though we're not asking for updated testing. We're just asking for, the level of testing is exactly the same. So reliability testing as we've seen should be consistent and the same thing with validity testing.

You might expect that as the

measure becomes more mature additional information will be available, but that's not a requirement. But the level should be at the absolute same, and this goes back to our conversation yesterday. It should be the same for maintenance measures and new measures coming forward.

DR. ASPLIN: So, for clarification, we should consider the whole of the material that was submitted for the packet that we reviewed along with the original testing.

And if an individual on the committee was going to determine that it does not meet this criterion that would mean that that person didn't feel like it crossed the bar originally, correct? That's what I'm interpreting your comments to mean.

MR. AMIN: Correct. However, the requirement of the committee doesn't mean that, there's no expectation the committee is holding the same bar as the prior committee.

I just want to make that clear.

Just because the prior committee may have felt that way doesn't mean that the current committee needs to feel that way. Everybody needs to evaluate what's in front of you compared to the criteria that's in front of you based on your own assessments of the criteria. It's obviously an objective evaluation, but there is some subjectivity to how you weight these particular components.

So I'm not trying to sway the committee in one direction or another, absolutely in no way. But I just want to make it clear that they should be held to the exact same standard of the criteria regardless of whether they're maintenance measures or not.

DR. LATTS: Was there other testing then that was originally submitted, I guess this is for NCQA, that was originally submitted that was not submitted as part of the resubmission?

MR. REHM: Ben, maybe you can

respond, but my recollection was that in the original testing in 2005, which you have, was the same testing we supplied two years ago when all of our, we have four other RRU measures that are NQF endorsed so that was the same.

MR. HAMLIN: Yes, it was. I mean the only additions were the reformatting to the measure testing formally supplied. Some of the additional analyses that we conducted in more recent years based on the annual submissions, but in a more limited fashion.

DR. ASPLIN: All right, we're going to move to Nancy. Before that just to follow-up to your question, Cheryl, from Ariel online said, medical policy varies by plan so the rejected claims volume could vary. I just wanted to note that for your question.

Nancy?

MS. GARRETT: So a question for NCQA about risk adjustment and socioeconomic status and sociodemographic factors. So you

provide evidence in the documentation that there are potentially disparities by race, by gender. I don't see anything specifically about socioeconomic factors, but I would imagine that's possible as well.

So you're risk adjusting in the models for gender, if I'm understanding this correctly, and technically that's the current position of NQF is that should be stratified for rather than risk adjusted for, although there's a committee looking at it right now and that's probably going to change.

So can you just talk through that a bit? It looks like also there is the possibility of stratifying at the health plan level by gender. That there's a way to report it separately even though you're risk adjusting for it. So can you talk about that a little more?

MR. HAMLIN: Yes. So the current measures are, the HCC risk adjustment approach basically predicts utilization and that does

take gender into account. However, we do ask the plans to -- I'm sorry. We do ask the plans to submit the age and gender cohorts because we do actually report out the benchmarks by age and gender cohorts.

So effectively the age and gender are taken into account in the risk adjustment approach, but we also then, we do report them out in a stratified fashion by risk cohort. So I don't know if that answers your question or not. We do not collect --

MS. GARRETT: So what does it mean to report in a stratified fashion if you've already adjusted for it?

MR. SAUNDERS: I'm sorry, Ben, let me jump in. We report the strata back to the plans, but when we're releasing information it is all at the plan level for the performance.

So like you said, it wouldn't make sense to have sort of the age group strata if you're adjusting for age, but it's a part of the math. We're still calculating all those

observed-to-expected at each of the cohort levels just as part of the nature of the math of working with the dataset, but is really reported out is the health plan level information.

MS. GARRETT: Okay. So I guess the question for the committee is whether people have any concerns about that. And then kind of my follow-up question is around other sociodemographic factors that are related to resource use, so things like race, ethnicity, language, education, income.

What are your thoughts on the relationship of that to these outcomes and whether the committee should consider recommending that the results be stratified in some way given that right now the NQF guidance doesn't allow for the risk adjustment up front?

So it's kind of a question -

MR. HAMLIN: We continually test the request data from health plans about their

consistency and completeness of their socioeconomic, SES, race and ethnicity factors, and then unfortunately it's too highly variable across plans right now for us to require it.

We've heard from some plans that they're actually actively not collecting that information for legal reasons, and so therefore there's problems there.

I agree that all of those factors could affect, you know, and do affect, probably, the resources used and the opportunities for resources used. However, there are limitations in the data for us to be able to actually stratify these measures by those factors at this time. And we are paying close attention to the current SES and sociodemographic risk group discussions.

DR. ASPLIN: Okay, I have Janis and Andrea and then Cheryl.

DR. ORLOWSKI: So I have two comments and I believe that they're both

directed at NQF. Since this is a re-up of something that has already been approved, I think it might be helpful for us to take a look at what we would expect to hear about the use of this during a three-year period of time.

We might look into what questions have been raised either to the developer or to NQF during the period of the time, if there's any evidence of its use within the medical community, any concerns that have been raised about the appropriateness.

So there's likely a way that we can track this and then require some update having to do with the use of the measure, the concerns that are raised with the measure during a period of time.

The second comment that I'd like to make goes back to both the comments that Bill and Lisa made regarding the goal. So yesterday we spoke about whether or not it is appropriate for us to make comments on should

the goal be higher, should the goal be low, you know, where are you headed. And I thought that those were very important comments yesterday.

And so as you take a look at this goal, and I recognize that we're not in the position to do this so again it's a comment for the committee to think about. As you come up with a goal, I think that there's three possibilities with this.

One is, you say there's no goal, you know, it's unachievable and let the market decide what happens. The second is to suggest that there is some ideal utilization that is appropriate, and I actually think that that would be very difficult. I think it would be very controversial, very difficult.

The third would be to take a subset of cases that are within this goal and then have an audited and peer reviewed assessment of the appropriateness. And I believe that you then begin to develop a group

of cases or an audited group within this that becomes an ideal or, you know, at least through what we currently believe is an ideal management.

And so again, I don't know that there's anything that we can do right now, but I do believe that it is appropriate for us to discuss how we would develop a goal and how we would develop recommendations for the appropriate use. Thanks.

DR. ASPLIN: Thank you. Andrea?

DR. GELZER: I just wanted to respond maybe to Nancy that we, you know, as a Medicaid managed care plan taking care of all the vulnerable populations, I think most Medicaid managed care plans do measure individual level race and ethnicity data. And I don't see why we couldn't do that with this measure, and similarly use geocodings, surname analysis to help fill in some of those gaps.

So I don't -- yes, go ahead.

MS. GARRETT: I think it's a great

point. And, you know, I don't think data limitations should be a reason for us not to make a recommendation, because making the recommendation can help improve data -

DR. GELZER: Right.

MS. GARRETT: -- collection and availability. And there are ways to do it on an aggregate basis, especially for health plan, using geographic analysis and things like that.

DR. GELZER: Right. And certainly if I'm submitting a measure, any measure, to NCQA I would do that analysis to the best of my ability, and I don't see why this would be any different than any other one. That said, I have a question to NCQA and the measure developers.

So for those of us on this committee who are struggling just a little bit with statistical analysis and validity and what is valid and what is not, why do you, I mean you're hearing lots of discussion here.

I would just ask you guys to make the case, why do you feel that this is valid? Why does this meet the bar?

MR. HAMLIN: It's kind of a loaded question.

DR. GELZER: I'm sorry. I'm sorry but, you know, again I -

MR. REHM: And if I can just set up, Ben, I'll let you roll with it, but great question, Andrea. And I also want to come back and talk, Nancy, about your question as well. So should I do that first and then come back to that?

You know, we developed a HEDIS measure that was just approved a couple years ago collecting race and ethnicity and language from health plans that report to us as a first step in achieving a future world, which is exactly where I think we all want to go and certainly where the IOM has told us we should be going.

In the Medicaid and Medicare

population there are different challenges around that data collection, but there are less barriers there than there are in the commercial population.

There are many regional barriers about whether that is a cool thing to report or not. Employers are very sensitive about this. Both Lisa and Andrea were both very involved in working on disparities reporting with health plans back in the day, if you will, and made great progress.

We're currently initiating a project, I can't really speak too much about it because it's not out there, but our committee on Performance Measurement identified this as one of the top cross-cutting issues, getting to this, and developing a plan around that.

And so I would just say we are actively engaged in trying to take the next step which is getting reporting. Would it start with RRU? Probably not. It'll probably

start with something straightforward.

No measure's straightforward. Breast cancer screening, cervical cancer, something out there in the space that where there's a quality gap and we want to understand more about that.

One of the distinctions about NCQA measures is that by and large we collect things on Medicaid and we have to think about that population and how we specify it. And we do obviously Medicare Advantage plans and commercial plans and divide commercial plans into two types, for better or worse, HMO and PPO.

So, you know, we are trying to bring that level of the data down and we have some tools available, and we think just measuring race and ethnicity and language of your membership even though it's not linked to a measure is an important first step.

But I think we are very interested in moving quickly towards the world that

you're trying to describe.

MS. GARRETT: And are you also considering socioeconomic status in those discussions?

MR. REHM: You know what, that is a, I'm not going to say it's outside our pay scale, it's that we really appreciate the work that NQF is doing in bringing together folks to take a look at that and come down with a recommendation.

And if that recommendation, you know, is X, then we're going to respond to that and think through that and understand the challenges and try to overcome them. That's our lifestyle. We overcome measurement barriers all the time.

Measures are not static. They're constantly being refined. The very measure we're looking at today doesn't even look anything like it was when we had kind of created it in 2005. It's changed a lot and it will continue to change.

So back to Andrea's question, and again I want to help set up Robert whose analysis, and Ben who helped develop the measure. If you take a look at the members on our Efficiency Measurement advisory panel, I think you'll see many of the people who are really bringing cost and quality into the national discussion.

I think just from a face validity perspective and from the fact that every time we put our measures out for change in the RRU, measures changed enough in its life to have gone through three or four public comments through NCQA's 30-day public comment period, we literally receive thousands of public comments.

And so there's been a lot of review. Even by the people who are burdened with calculating the measure, this is not a simple walk in the park. And they come back to us and say, it's hard. It's a pain. You've made it easier. You're providing us

valuable information back.

This is not a one-directional measure. This is an interesting measure. We shoot as much information back to the plans, probably more than they're actually providing to us because we can help interpret it and give them national benchmarks back.

And so I think that from a validity perspective, which is the question you asked, Andrea, I think at one level, at a fairly high altitude level, we believe this is living in the true marketplace. If the marketplace didn't like it, didn't find it valid, they would have rejected it.

But that's just the high - Ben, do you want to pick up on that?

MR. HAMLIN: I just wanted to add to the fact that, you know, I mean, these measures again were sort of initially conceptualized in 2005 to address a very specific need, and that was the fact that there really was not a lot of information

about value.

I mean, I know that healthcare costs continue to be in the national discourse, and again we've heard about the limitations about being able to do that. These measures are constantly evolving. They were not even publicly accorded or the information was not available back in the public sphere for several years.

I believe it was 2009 was when this measure was first available for public reporting, so it was, you know, a number of years of development and then a number of years of refinement. So even as a result of the first evaluation by NQF for this measure we made a number of fairly significant changes based on that feedback.

So the biggest one was in our list of inclusions where the committees, I don't remember which level of committee, the steering committee made a number of comments about their concern that we were excluding at

that time ESRD patients who they felt were -- or, I'm sorry, patients in this cohort who were identified with ESRD, and they felt, you know, for the diabetes and cardiovascular measure that was actually a pretty critical component to be included.

And so we did go back and do some additional testing and some research in our large stock and database and determined that, you know, that was, in fact, correct. And so therefore we worked out a way to include that in the iteration of the measure.

We are constantly looking to include additional services. You know, we have to ensure that the services that are being priced and are being included in the measure through our standard pricing tables are, in fact, relatively reliably priced and fair, if you will, to assign a price to that service at the code level.

And we have been able to in the last few years to add additional service

categories, the diagnostic laboratory and diagnostic imaging, after some additional testing that was, you know, found that we could, in fact, price out the vast majority. I don't want to say all of those services.

So again, you know, in receiving feedback we are constantly making updates to the measure to make them more relevant to the audience that they are intended for which is both the health plans to help them identify their performance, if you will, against their peers and presenting to the consumers, and including consumers, employers and government entities such as the state officials who are interested in the report cards, and others that, you know, they have the level of detail that they need.

And so, you know, again we're constantly revising, we're constantly re-looking at the measures and looking at the data that comes in to make sure, that is, the measures still work for that specific

environment for which they were developed.

DR. ASPLIN: Thank you, Ben. Tom?

DR. TSANG: I have two questions for the NCQA. Number one is, can you elaborate a little bit more about the conversion of ICD-9 to ICD-10 and the impact? I know there's one statement here in the TEP about annual updates of the value set and implementation of additional codes would address this issue, but I'm just wondering, you know, the expansion of the number of codes from ICD-9 to ICD-10, it's hugely significant. And I'm just wondering if any of these validity testing will be impacted by this conversion. That's the first thing.

And then the second question is really refinement of this measure and the improvement of this measure, I'm wondering if there's any distinction or differentiation between correlation of these measures with claim space quality measures versus this measure correlating with eMeasures, clinical

eMeasures, and have you seen any difference in those types of associations?

MR. HAMLIN: So to answer your first question, NCQA for all HEDIS measures has published the ICD-10 codes that are mapped from the ICD-9 codes currently in the measure. So the first part of that was to sort of try and do the mapping and review.

We have an expert coding panel that looked at that and performed that and those ICD-10 codes were published alongside those. That being said, we do an annual review of our code list that identify all these conditions for these measures and that's based on both public feedback, expert feedback, new codes being available, old codes being retired and so on and so forth.

And so there's a whole process for that. I would expect that, you know, given the comments from the clinical committee that, you know, with the increased specificity of ICD-10 and the additional things that might

perhaps be included under our definition of cardiovascular conditions could be expanded, I would agree.

And I think we're going to closely look at that as our comment, you know, I think in that response said that, you know, we're very interested in making sure that we're including the appropriate eligible population.

We do have several sort of small projects with individual organizations looking at dual coding right now. So some organizations have already started dual coding, you know, to prepare for October of this year, and we're working closely with them to try and understand the effect it will have on HEDIS measures. Unfortunately only time truly will tell what the actual effect will be, but I believe it's going to affect the vast majority of people. Well, I know it will affect the vast majority of people.

You know, we're hoping that our caution in including new codes or in watching

this transition will pay out and that it's not going to completely undo everything we've developed in our latest portfolio.

With regard to eMeasures, NCQA is intimately involved with both the measure development and many of the meaningful use programs and contracts both on the software vendor certification side and also in eMeasure development.

Currently the eMeasure specifications and reporting programs are not mature enough for us to truly be comfortable with the data that's coming in. You know, again the issues of whether we're reflecting national performance or whether it's a data submission issue, again these measures are not in a full reporting program that's audited and validated. You know, the stream of data has not been validated and audited yet.

So we are very intimately involved in making sure that the eMeasure specification process is adhering to our standards of

transparency and scientific acceptability and measurement reliability. However, the results from eMeasure reporting are not significantly mature for us to be able to look at the correlations between measures derived directly from clinical data to resource use at this time.

I am sure that some systems who have much more access to internal data and have much more mature EMR platforms perhaps are looking at that. But at the NCQA level for the national reporting program we're not able to do that at this time.

MR. REHM: Paul, just to supplement Ben's comments. Just like we do for all of our measures, we took three years to essentially transition all of our coding to ICD-10.

So we have about 100 measures out there and we split them into three little groups and spent a year putting them out. And at each different cycle we put all of those

new code sets out for public comment. So I just wanted you to be aware of that process.

DR. ASPLIN: Good. I'm going to go to Dolores and then we'll have Taroon and Ashlie make comments, and then we're going to discuss the algorithm reliability, provided we don't have other questions.

MS. YANAGIHARA: It's a quick question for NQF, actually. Did these guidelines for reliability and validity actually exist when we endorsed the measure a couple of years ago or are these new since then? I can't remember.

MR. AMIN: So the guidelines for evaluating reliability and validity existed. The criteria for testing hasn't changed. What, this algorithm has been developed by our lead methodologist at NQF to help committees be more standardized in their application of the criteria.

I will also say, as you remember, that was the first sort of cost and resource

use effort so, you know, I don't know how familiar everybody was with the criteria so that might be another input to consider.

MS. WILBON: I was just going to add one more thing to Tom's question about the ICD-10 codes. We don't currently have any requirements around testing of the measure for ICD-10 yet because of the limitations of the data available to actually run measures on ICD-10 data for those organizations that do have the ability to dual code, which is not very many have the resources to do that.

A lot of people don't have access to the data to actually do the testing, so that is not a requirement that we have yet, just for the ICD-9. We request that they submit them, but the testing, we haven't yet made that a requirement yet.

DR. ASPLIN: Taroon, can you make your comment and then actually walk us through the whole algorithm?

MR. AMIN: Yes, actually that was

kind of where I was going to go as well. And actually I wanted to frame it a little bit as a question for the committee because I know it was subtly addressed a little bit through the lead discussants, and I know Cheryl sort of made some comments to this effect too, but I just wanted to be really clear about this.

I guess the question I have for the committee is really, I think, you know, there's some questions about Number 1 here that the TEP has raised. But even Number 2, I guess, in particular, can we have a discussion about to the extent to which there is empirical reliability testing that was submitted in the attachment?

I know many of you have reviewed all of the attachments, not just the testing attachment, but what I've heard from many members of the committee is that this is sort of, there's a lot of descriptive information and process information.

But I'm trying to understand the

level of empirical reliability testing because that will have a very clear impact to where we land on the algorithm.

MR. HAMLIN: Taroon, I may make a comment to NQF that if you would like a lot of detail on testing you should not limit the testing form to 20 pages.

MR. AMIN: I appreciate that comment, Ben. Thank you.

DR. WEINTRAUB: Taroon, I think you summarized it well. It's mostly descriptive data but not a lot of real reliability testing.

DR. ASPLIN: Lisa?

DR. LATTS: So Ben, in response to that comment, then are there other data that you have that were not shared as part of the packet?

MR. HAMLIN: Outside of the attachments that we can provide, I mean, we do have, like I said, we do an annual analysis that I believe we used as our testing, as a

good part of our reliability testing information for the original submission.

And those are fairly extensive correlations of the different categories both to each other and to the quality side and also the outlier analyses and the plan quadrant shifting analyses that we do.

DR. LATTS: Thank you.

DR. ASPLIN: Is there more specificity from the committee about what you exactly would have liked to have seen either from the original testing or the annual testing of the measure's performance that Ben just described? Yes, and kind of a qualitative description of it.

So from the original patient level data reliability testing done in 2005 or 2008 we got summary information, but if there's a lack of satisfaction what specifically, especially those that have the methodological chops here, what are you looking for?

Cheryl?

MS. DAMBERG: So I'm looking at, I guess it's Page 32 of the documentation and it talks about, this is in the reliability section. It says an indicator of plan stability over time is quartile movement of O/E ratios with significant shifts having implications about plan performance in terms of resource use.

And then it says, for comparative purposes plans that move less than one quartile are considered stable. So it's very descriptive and it doesn't give us a sense of, you know, when they've looked at the data how much shifting around is there.

And I think a table here that would help people see how much movement there is since, you know, this is their primary means of demonstrating they have reliability. So I think it's more the quantitative piece seems to be missing here.

MR. HAMLIN: Are you interested in the number of plans shifting or the magnitude

of each plan's shift?

MS. DAMBERG: Both. Yes, I mean it would be helpful to know, you know, maybe this isn't the right analogy here, but this is the type of work that I do.

When I'm looking at a performance measure, I look to see how many rank positions any given provider moves depending on what I'm doing in the analysis. And so if I see big shifts, you know, that's more troublesome than if I see shifts of like one or two rank positions.

MR. REHM: Ben, if you can help me find in the annual report, we provide to the committee on Performance Measurement an annual report on RRU. The last annual report was about 80 pages. I can just show you the graphics here. This is the data that doesn't fit into 20 pages.

But Ben, which table includes the quartile shift that Cheryl was asking about?

MR. HAMLIN: Well, again we only

provide the number of plans that shifted, we don't actually provide the magnitude. And I think that was the first part of her question.

MR. REHM: Right. Actually do you have access to this? Taroon, was this included in the packet that we sent you? Okay, well, I could try to read the table here. It shows the percentage of plans here. You can validate this, the percentage of plans that shift.

I mean, clearly we can provide you this voluminous information. I think what it tells you, it may not tell you exactly what you're looking for but it answers the question, are we exploring the detail each year of the performance on the measure.

Are we looking at, in this case, quartile shifts? What percentage of plans by each plan type are moving and what's the extent of that movement? Yes, we can answer that question.

MS. DAMBERG: That's great. That

would be really helpful. Because I think part of what we're trying to judge is to what extent, where plans get classified is purely random based on sort of the signal that's in the estimate. But I suspect it's not, and I suspect there's some stability but we weren't able to judge that.

DR. ASPLIN: Thank you. Andy?

MR. RYAN: So Brent, just to respond to your question. I think in the last several submissions we've seen, and of the reliability coefficients we've seen this split-half correlation yesterday, and we saw kind of shifts in groupings over time that Cheryl just described. So I think some combination of those things to provide evidence of reliability is what the committee would be looking for.

DR. ASPLIN: All right. Yes, go ahead, Bob.

MR. SAUNDERS: I think one element of that, I mean, so we definitely have the

beta binomial information. The challenge we sort of run into on that is that with the risk-adjusted measures that once we take out all of those explanatory factors of the risk profile of the patient it sucks away the variation, and so the reliability kind of naturally suffers on that. And so we don't think that the beta binomial is sort of the right choice for that.

I think where we would lean towards as sort of thinking of we definitely have the information about the proportion of plans that are moving and how stable is performance and so I believe we can provide those percentages.

I think the other way that we think about this is sort of what is a meaningful shift and the observed-to-expected ratio and thinking about sort of a criterion based approach to can you distinguish the folks that are above or below some threshold as a way of thinking about that. But I don't

believe that that is in the testing materials supplemental.

MR. AMIN: Okay, so Brent, to your issue around the algorithm, I think the basic question that we need to be focused on here is effectively 2 and 3 of this, essentially to the level of empirical testing we believe that's been submitted. And we can think about it broadly, you know, to recognize the concern around the 20-page limitation of the information that was presented.

And in our actual NQF submission form, the developers have provided their technical appendix which I know at least, I mean, I'm sure the committee has reviewed in particularly, the methodology folks that have spoken on this particular issue as well have certainly thoroughly reviewed.

So that's the complete information that we've gotten from the developer. So as you're making this decision you really should be assessing essentially that question at this

point for reliability.

MR. SAUNDERS: Just real quickly, they were able to pull up out of the table it is between 89 and 94 percent depending on which metric in the cost categories are in the same quartile or move one quartile up or down. There's not folks moving from the best group to the worst group and vice versa.

DR. ASPLIN: Very good. Jack?

MR. NEEDLEMAN: I'm trying to think about the issue and the challenge of testing reliability in the NCQA context. And, you know, what the reliability measure is fundamentally about is about the stability of the rankings in the face of data jitters.

And you do a lot of testing out of your data warehouse at the patient level and that'll tell you something about the stability of the measure around, you know, how much variance you get once you begin pulling subsamples from that.

But at the plan level you don't

have the patient level data so you can't sort of redo the plan. So I'm just wondering, you know, thinking down the road, thinking about going back to your data warehouse and creating some synthetic plans out of that and then playing around with subset analysis, reliability testing around your synthetic plans and seeing how much the relative rankings shift and how much the score shift might provide the kinds of information that people here are asking for.

The year-to-year variations are affected not only by the changes in the patients, which is what we're trying to capture with the reliability, but also by the changes that the plans are doing in trying to improve and we can't separate that.

Some stability is expected because we expect plans to move slowly, and basically the numbers you just gave us suggested that that's what we're seeing. Small movements consistent with change taking place slowly.

So it's comforting but doesn't quite feel sufficient to really meet what's in this rubric here that's on the screen. But I think that's one of the challenges that we have as a committee thinking about the limitations and the way NCQA is collecting its data and doing its analysis at the plan level.

MR. HAMLIN: One of the other limitations to address that very point is I don't think we would expect every plan, you know, our assessment of quartile shift is at a role of level of all individual data.

I mean, I would expect that plans would not focus on each, so we provide information back to the plans at each of the individual service category levels and their comparison to their peers on each of those service categories.

And I would expect a plan to probably focus on certain areas and not be able to do everything at once overall because that would just be unrealistic and probably,

you know, cost prohibitive.

So I do expect that plans are, year-to-year-to-year, looking at each of these service categories and comparing themselves to their plans and seeing how they compare, you know, to both their immediate peers and to the mean, if you will, how far away they are from, you know, that standardized mean.

But again, I think that's the detail at which the plans then take this template and go back to the, you know, to do their own internal analyses, whether -- but that's again beyond the level of what we can actually do because of the level of information that we receive.

So our plan database does actually include actual plan data, so the database is updated with, I believe it's about 60 or 70 health plans at the moment that gets, you know, occasionally updated using the actual plan data.

So we could theoretically retest all of

the assumptions we did in the initial testing for appropriateness and validity for, you know, things like risk adjustment to this approach, however, that also is extremely costly and given a lack of any other reason to do so, you know, at this time, I think, again with ICD-10 rollout we will definitely want to go back and do that.

But I think until that data and that experience from ICD-10 is mature enough for us to be able to do that and in order for our database to be populated with that information it would be probably against our own best interests to do so.

DR. ASPLIN: Bob?

MR. REHM: Thanks. Jack, appreciate your comments. In some ways you're asking the question, is this measure different? And algorithm aside it could be that it is different. But I do want to provide the committee kind of a trajectory.

Out for public comment as we speak

are new standards, new measures for accreditation in programs. We accredit a thousand health plans around the country. And in that discussion where we're really revitalizing that measure matrix, but part of that discussion was thinking about RRU.

And I think the challenge that we faced as currently as reported and how we're handling the data, the ability to -- oh, just so you know, that when we accredit plans using measures it's 50 percent of their accreditation score and there's thresholds, 90 percentile, 10th percentile, 75th, mean, median and all that stuff.

And we created benchmarks to apply appropriate credit for better performance or lower performance, and a key criteria for that is being able to differentiate the very performance you're asking us to prove to you that we can differentiate.

And so I think that, now those measures aren't in that public comment

because, one, we felt at a measurement level that we need to let the ICD-10 stuff sort out first. We don't want to be testing stuff before that gets played out because that's important.

Second, we feel that there's techniques that we can apply, and whether it's a virtual or a synthetic health plan conglomeration where we can redo this stuff, which is something we thought of and that the trajectory again is to get to this ability to benchmark, when we get to the ability to benchmark, satisfying the algorithm or any other kind of criterion is not going to be hard at all because we'll have had to prove to the field that we can differentiate that.

I think in so many ways this measure is a -- speaking of signal-to-noise, this is a measure that we're trying to signal that value matters. We were early in the field on this. The NCQA investment on this measure, not that that matters, was over $1

million. That's our own money, not anybody else's. It's not a profit center.

So, you know, we're trying to refine the measure all the time is why we do this exhaustive annual report. Look at correlations, try to uncover if there's anything interesting going on that we can latch onto that help us make it even a better measure.

So I do think that if the measure were to pass or go through and we're back in three years that you would probably be seeing something more refined than the measure you saw two years ago. So I guess what I'm talking about is there's a trend here and we're interested in proving the measure and proving to you what you would like to know. And I appreciate the challenge.

DR. ASPLIN: Thank you. Janis?

DR. ORLOWSKI: So I understand that this is a relative measure. My concern is that the entire market could have dropped,

I know it hasn't, but could have dropped by 50 percent of actual dollars. And what we would be looking at in this measure is the plans would still look at each other in comparison to are they spending less than other plans.

So it tells us where the plans are relative to each other, but gives us no information about a relative shift within the market.

MR. REHM: So to the extent that we're trapped, if you will, by standardized pricing to the extent that this is driven by RBSs or whatever, those things may be dipping but they're probably dipping slower than the actual market may be dipping in a particular area around cost.

You're correct. In terms of the health plan taking those data and then basically tossing in their own cost data, then that's meaningful. This is about a measure that's not operating just in this space.

It's a measure that's operating in

this space and then moving down as Dolores was talking about, their review of using the measure at the provider level or, you know, obviously ACOs are an area of great interest and I would imagine would be working on that as well.

So I think the U.S. knows that the cost trend has been slowed. You know, we know enough about the macroeconomics of this to appreciate that. Some markets have great transparency about what's going on. Others don't. So it's a varied issue there.

So I mean, I think your point is well taken that the storyline from almost a policy level maybe the measure doesn't tell you as much as you'd like to know, but at a micro level it's probably more informative.

DR. ORLOWSKI: So during the last three years, PCIs have gone to outpatient almost, you know, 90 percent of them. Chest pain, the first 24 hours is an observation status. So there has been a dramatic shift in

what is paid for cardiovascular services.

And again, I think it's just a point that what we're seeing is plans relative to each other, but you don't see this shift that has occurred because you're not taking it. So I bring this up because again I think the measure is what it is.

We understand that it's a relative measure. It doesn't give us information about a change in the market which has occurred in cardiovascular disease, and again I think it raises the issue of what the goal is.

MR. HAMLIN: So I think I would like to make one point there. I think you would -- so the standard pricing tables actually are reflective of whether the service was offered in an inpatient or outpatient basis, and so there's an adjustment based on the coding practices of that, you know, that outpatient and inpatient services would be effectively fairly priced if in a standard manner.

The measure itself also is broken down by inpatient and outpatient, so a plan would see a reflection of shift in the market from inpatient to outpatient services for very specific areas, but certainly in procedures, evaluation and management and so on and so forth.

Those are individually reported service categories within the RE measure. If the entire market shifted in that regard, yes, you probably would not see that shift in the measure itself. However, you know, on a national or even on an HHS regional basis, I think that shift may have been a bit staggered or a little bit longer, and I believe the measures would probably pick those up especially at the plan level. So I just wanted to offer that additional information.

Also I think on the policy context, you know, I think over the last few years the fact that an increasing number of plans choose to collect and report the RRU to

NCQA is significant in the idea that these measures are, in fact, valuable to somebody.

DR. ASPLIN: Thank you. Lisa, I'm going to give you the last word before we move forward with a vote in reliability.

MR. WILLIAMSON: So two comments. One is sort of in response to Janis's question. I'm just a little confused by it because it's not, what your question was getting to, to my mind, is not what this measure is at all designed to do.

And, in fact, we have lots and lots of things that do what you were asking, not the least of which is premiums which are based on overall cost of healthcare, and now especially in a post-ACA world we're limited to an MLR of 85 percent. So we know exactly what's going on with the overall healthcare cost.

So to my mind, what this measure is actually designed to do is something totally different. We have actually a far

better handle on overall healthcare cost from many, many perspectives than we do on relative resources. So your question actually confuses me a bit.

My question, my original question actually was more towards NQF again. If this committee were to say it's not passing reliability and based on the testing essentially that was done before the original measure, does that reflect at all on the original process?

And would it, I guess, and I'm sure there's a more diplomatic way to put this, but would it call into question the inter-rater reliability, essentially, of the process?

DR. BURSTIN: It's an excellent question. There's no way for us to always have a sense of our own inter-rater reliability of course. I do think, you know, the timing of the original report, and I know the original work was done after our testing

task force was put into place where we had a new report, we had raised the bar. It was not very long after it.

So I think some of this is, you know, based on guidance from many of you we have continued to raise the bar. I think, you know, as I mentioned yesterday is that we recognize it as a real challenge for measure developers to continue to test measures in use.

And I think you've heard, you know, from NCQA, there's a fair amount of on-the-line surveillance in evaluation of the measure. But I think I would more so just focus in on sort of where we are right now.

I don't know whether there's additional information that could be brought to bear from NCQA that might influence that decision, but certainly we afforded that opportunity.

MS. WILBON: I would just add, Lisa, that it was a different group of people,

and considering that there are many new members on this committee and this is their first time seeing the measure that I'm not sure if you can really compare, because it's not the same rater rating it twice. And so to that extent is one of the reasons why we're implementing standing committees so that over time we have that consistency.

So I would rather us, you know, not, I wouldn't worry about the previous committee at this point. Let's just consider this kind of ground zero going forward. If this measure were to come back it would be the same. You guys would still be the standing committee.

So from that point I think we have a better idea of whether or not this, you know, the measures we're implementing to try to maintain consistency are really working. So I'll just add that.

DR. LATTS: Yes, it's hard for the developers, because, you know, from their

perspective, oh, this is what worked when we brought it before the committee the last time. And so it would seem to be a challenge from a developer perspective to know what to bring forth to the committee to meet the committee's needs, sort of a priority.

DR. BURSTIN: And that's been the work behind now having a measure developer guidebook, trying to put these algorithms into place. It's, you know, I will acknowledge that it's certainly a work in progress for all of us.

DR. ASPLIN: Nancy?

MS. GARRETT: So earlier I made this proposal that we might want to consider making a recommendation that this measure be stratified by sociodemographic characteristics. It sounded like Andrea had some interest in that, but I'd like to know what the committee thinks about it, and then is this the right section to be talking about that or is that in another place? Is that

under usability or --

DR. ASPLIN: It feels like a usability issue. You know, recommendations on how the measure be used, I think that would be the section for that discussion.

Okay, there's only one way to find out what's going to happen, because I can't read what's going to happen. So let's move ahead with the vote.

MR. WILLIAMSON: We will now vote on subcriteria 2a for reliability. We have four options, high, moderate, low or insufficient, and you may -

We have all the votes. Yes, so we changed the interim view so we could see all the votes. It looks like we have 12, not 12, 18 moderate, we have two low and three insufficient. The measure passes reliability.

DR. ASPLIN: And we are going to take a break. We'll resume at 11:00. I'm not going to say another word.

(Whereupon, the foregoing matter

went off the record at 10:48 a.m. and went back on the record at 11:03 a.m.)

MR. WILLIAMSON: Before we get started I'm going to read off the votes just to make sure we're all clear. Hi everyone. Before we get started again I want to make sure that we read off the tallies that we've voted so far just to make sure we combat those technical issues we had earlier.

So 1a high priority we had 20 high, two moderate, zero low and zero insufficient. For opportunity for improvement we had seven high, 13 moderate, two low and one insufficient. For 1c for measure intent we had 17 high, six moderate, zero low and zero insufficient.

For overall importance we had 12 high, ten moderate, one low and zero insufficient. And for reliability we had zero high, 18 moderate, two low and three insufficient.

DR. ASPLIN: Very good. And then

before we move on to validity, I guess to Ben and the developers, one question I would have is since similar measures are coming this committee's way and the same questions would come up, a clarifying question. Would we expect new testing prior to ICD-10 being in place? I believe the answer to that question was no.

And a related question would be if the answer is no could we get a more complete submission of the original testing and detail that goes beyond the 20-page limit so that we could dive into that prior to the subsequent measures being discussed at future meetings?

MR. HAMLIN: I think the answer to your first question is comprehensive testing prior to ICD-10 would be no. And then the answer to your second question is we are certainly happy to provide additional documentation that would fill in the back story if you would like to request it.

DR. LATTS: This is Lisa. Could

we also get that documentation for this measure to this committee? Just the report that you were reading off of.

MR. REHM: Sure. We can provide you the annual report, and I think you probably already have the original field test.

You know, I think you raise a good point. Some of this is packaging. In many ways this measure is context-driven and I think you've captured that and appreciated and understood that the submission form, the way it's broken out it's hard to tell a story.

And we can certainly go back and ask ourselves, okay, it's not exactly what we thought would fit there but let's go ahead and drop in the stability data, or we did assessment of some of the risk adjustment models, et cetera.

And so we'll do our best to package that a little bit better for you to make it easier. We do have three more measures coming up in this space, so happy to

do that.

MS. WILBON: Also I would just suggest we do offer technical assistance to all developers, and it would be really great to have an opportunity to, we do offer to help developers kind of help package that material. And so we would offer NCQA the opportunity to meet before hand to make sure that we have the most succinct information in there for the committee before it's submitted.

DR. ASPLIN: Nancy, do you have a comment on reliability or are you just ready to roll on validity? Okay, good. So, you know, there's always danger in interpreting votes because there's probably 23 interpretations, well, there are going to be 23 interpretations of the vote. We're not going to go around the room and describe them.

But so my takeaway from the last vote is not that there's a great enthusiasm, but rather that there's some degree of trust that some of the testing had been done, and

also trust in the original committee's decision looking in more detail at some of the reports.

And so I guess my read from that is that since some very similar measures are coming our way that we would request that NCQA not take the vote as a sign that we can have the same discussion three or four more times, and hopefully we could get more information on the table so we can have a greater degree of comfort.

So it's just an editorial comment. Let's move forward with validity, and I'd like to start first to see --

MR. AMIN: Brent, do you mind if I make a few comments on that? So a few additional comments that I'll make to the committee. The first is as we put these votes out to public and member comments, as we're thinking about internal consistency between this group and the prior, be very mindful of the internal consistency between this measure

and the measures we reviewed yesterday. So that's the first point I would just make.

The second is I would like to have a conversation about the empirical testing. If we go back to the algorithm, the only way that this measure could have been rated moderate is if the committee felt that there was empirical testing.

The measure developers did present this information. It would be an expectation at least during the comment call that we'll review this information again in terms of actually having committee looking at this empirical testing. Because based on our prior conversation it didn't appear that there was empirical testing that the committee reviewed.

So I'm assuming that the committee was basing the empirical testing requirement based on what was given verbally by the developers by what's in their annual report.

The third thing which was sort of highlighted and I'm just going to raise it

just to keep in mind is, you know, the requirements that NQF puts together in terms of the submission form limits, in terms of page numbers, the format that NQF puts this in is to ease the interpretability for the committee to ensure that there's standardization for the committee's time and ensure that there's consistency across measure developers.

So when we have conversation, first of all, it's not within the committee's sort of authority to allow for additional, you know, more than 20 pages, but it is to have the information in a succinct way for interpretation and evaluation by the committee.

So we could take back the recommendations of the committee if they feel that there was not information or not enough space for developers to put this information in.

But the way that we've set up the

submission form is standardized across all of our measures and for our measure developers, so we want to be respectful of sort of if we're changing the requirements those requirements have to be rolled out, and those are understood and agreed upon by all measure developers in terms of the submission form and our criteria.

And this will be part of our larger discussion, but since it's clearly related to our prior conversation I just want to make this really clear. A few members came up to me during the break with the concern that are we really, actually using this algorithm in our decision making and, you know, are people extrapolating their own opinions about what needs to be done or the importance of some of these components?

And I think, you know, that's a reasonable consideration that the committee should discuss amongst themselves, and if there's a concern about any of the level of

the criteria also remember that those criteria are applied across all measures, not just measures across this committee but also across NQF.

Clearly each person will make their decision about how well these criteria are met based on their own expert judgment, however, particularly in this section of the evaluation it's intended to be much more objective.

So I would like to put all those topics on the table before we get to validity, and this is not just a conversation limited to this measure by any means or this measure developer by any means. But our role, increasing role of staff is to make sure that there's consistency and that we sort of have this break period to say, okay, are we comfortable with where we are particularly when members of the committee are not necessarily comfortable with where we landed.

And I think you should have that

opportunity to have a discussion with each other about that topic.

DR. ASPLIN: Lisa and then Andy and then Cheryl.

DR. LATTS: I want to comment on Taroon's -- yes.

DR. ASPLIN: Yes, I guess we're having another conversation first.

DR. LATTS: So I guess my question based on that, Taroon, is that we've seen three measures in front of the committee for this phase, none of which had the empirical reliability testing that the algorithm asked for.

So I guess my question is either there's a disconnect between what the developers are being asked for and what we're being asked to evaluate on, or what we're asking for is not -- and I guess, I don't know. Maybe we should ask Yale and NCQA to comment on this, but why are we not getting them what is being asked for? Because what

are we supposed to do as a committee?

I mean based on the algorithm then, none of the three measures that we've seen are, quote, unquote, high quality good measures, and then what are we all wasting our time here for, I guess, is my question.

MR. AMIN: Okay, so there's a few different components there, and I think, you know, the concern about high and moderate, I think, is a level of interpretation.

But there's a particular methodology, I think, and I don't know that, I mean I can't speak to the committee, but there were differences of opinion about how much reliability testing was done between the two measures. So I mean that would be difficult for me to say.

But I will say that to the extent there should be consistency around what we're requiring in terms of methodology, and the reason why that is is because we don't want to set a bar in which some measure developers are

spending a tremendous amount of resources.

Again, this is not describing anyone in particular, I'm just saying we don't want to set a bar which some measure developers are spending significant amount of resources in terms of methodology and statistical support and then others are not, and there's, you know, irregular sort of application of criteria.

Now if there's a concern about the criteria, meaning that, look, reliability testing is just not able to be done in cost and resource use measures using administrative claims data and we believe that that may be the case or the current state of affairs in this measurement domain, I think we need to state that and apply that consistently across all the measures that we're seeing here in this project and going forward.

And try to, I mean obviously that would raise another level of concern that we should say, well, what are we going to do to

address that globally, and maybe we need to put together a panel to give guidance around this whole topic or reliability testing, working with our developer colleagues to understand the current state and the challenges that there are of achieving the actual criteria.

But those are two different realities, and we just need to be really clear and transparent about what we're doing. Because what will end up happening through this process is that you go through public comment period and developers will feel that, you know, if it's not applied consistently the committee will have to address that through the comment period.

And it's much more difficult to do that once you have numeric values here that we're supposedly using the criteria and the algorithm to decide how we're making our decisions. So if it really is the issue that the criteria is too high of a bar then let's

have that conversation, not now but in our afternoon session, and understand what to do about it.

And we're totally open to that. We can remove criteria if we feel it's too high of a bar to, it's reducing innovation in the field or it's redundant with other measure developer processes or things of that nature. We're certainly open to that conversation.

DR. ASPLIN: Very good. I might not have gotten the order exactly correct, but I've got everybody names that's got the card up.

Janis?

DR. ORLOWSKI: So I recognize that we had an opportunity to speak with the developers on telephone conversations earlier in the process, but what I would say is that the conversations that we had yesterday and today have been very rich conversations with the developers. And they have influenced in several ways how you view the data that has

been provided.

And I wonder if there wouldn't be a more appropriate place for this discussion rather than five minutes before the vote. And so what I would say is that part of the discussions and part of the votes yesterday and today, I think, in a large part were affected by information that was presented immediately before the vote.

MR. AMIN: I know we're taking up a little bit of time but this is really important for our, this is broadly important. Because one of the questions that we're still struggling through with NQF is that our typical approach prior to this phase of work and our improvement work that, you know, many of the folks here at the table and just broadly have helped us think through is that the submission form is what you're evaluating. And that is what goes out to the public, that's what the public has reviewed prior to this deliberation.

And as you go forward, you know, the information that's presented what we've tried to capture and make sure that that gets distributed as well, but it's obviously challenging for the members to make comments about the measures if all the information's not in the submission form.

And so what we'll need to think through is the fact that the process that we have set up right now is supposed to be an objective evaluation of the information that's submitted in the actual submission form.

Now to the extent that there are additional questions and there's additional data that the committee wants to see that could be addressed, obviously we're not saying, you know, use blinders and that's not relevant here.

But also I just want you to be aware that the committee's deliberations are part of a larger conversation that the membership and public is part of, and to the

extent that we can that information needs to be transparent and the main transparency vehicle we have is the submission forms.

And that's why we ask the measure developers, that's why we ask you to make it very clear how the decisions are being made.

DR. BURSTIN: Just to build on that one second, just -- sorry.

MS. WALKER: I asked a direct question to that effect yesterday and you had indicated that we are supposed to use all the available information, which I did and I assume everybody did.

Now I would say that on that particular question during the webinar call, I and others on the phone had explicitly asked the developer to provide that information because we felt that we needed it to assess that particular question.

And having received that information, you know, it made the measure look a lot more favorable. And without that

data, the vote, at least my vote, would have gone a different way.

Now I think it also behooves the committee, I mean if we can incorporate these additional data, I think it behooves the committee to ask for that additional information at these webinar calls. Because it sounds like there are data that the committee would have liked to see, and it sounds like the developer has some, maybe not all but some of that information, and if that had been conveyed earlier in the process I think that that would have been more helpful to this conversation.

So that was my first comment. Can I make my second comment? So my second comment has to do with using the algorithm and in responding to what you were saying. Now I'm not a data statistical heavyweight, but listening to the conversation it sounded like what would be acceptable to a reliability test would be if there was more stability in the

rank ordering from year to year.

And the developer described what they had done which is didn't provide the actual values from that data. So in reading this algorithm, when you say answer no if it's only descriptive statistics, that to me is not descriptive statistics. That's more than descriptive statistics. They actually did the analysis. They just described the results rather than presented the results.

So I think it's important for us to understand that distinction and at least that explains how I voted.

DR. ASPLIN: All right, I have Cheryl, Andy, Nancy and Jack.

MS. DAMBERG: I think this is a helpful discussion and I would say I have to confess. I haven't looked at the measure submission form and all the instructions in detail, but it strikes me that given that this seems to be an ongoing challenge for the committee to review the materials of the

measure developers.

Maybe there's an opportunity to here to provide them some examples of what a good reliability and validity, you know, written section would look like. And I think largely what's missing in that section are results, you know, data so that people can judge.

And so I think you could dummy up some examples or maybe draw from some better submissions where people have produced that kind of information, because I'm sort of reminded of when people put together methods papers for scientific journals around measure testing they are essentially writing these two sections of the measure submission form and they're doing it in a very digested way because the journals have, you know, word count limits.

So I think if they can work toward that kind of model I think that might be helpful.

MS. WILBON: So we actually do have that, Cheryl. We call it the "What Good Looks Like" document. And we did actually, for those of you that remember, we went through that document with you guys on one of our calls, it was last year probably sometime, and we have that on our website.

And so we have been trying to make all the developers aware of that and there are several examples in there for different types of tests, inter-rater reliability, face validity of how to display the information in the submission form, the types of information we're looking for.

So it's out there and, you know, the degree to which developers are applying that in their practice and putting the submission form together, I agree there's still a disconnect there and we're doing what we can to try to work with developers again before the submission to make sure that, you know, they're doing that.

Unfortunately at the time of the submission we have a very short turnaround time frame, and if we don't have that opportunity before we don't always have time to do a lot of back and forth after the submission deadline to make sure that their submission is kind of as tight as it could be. So I think that's the challenge that we're facing.

MS. DAMBERG: So does NQF, I mean I'm not necessarily trying to put more power in your hands, but do you have the ability to reject a submission if it doesn't have that kind of information? You know, it's kind of like an incomplete college application, like you didn't do the college essay, so, you know.

DR. BURSTIN: Yes. This has been an interesting issue of when we feel comfortable having staff make an assessment of completeness versus, you know, if boxes are left out, sure.

But if there's information in there and

it requires a real qualitative assessment, we've been increasingly doing more of that in-house but we haven't at least allowed that information to flow to the committees to ensure that you have a chance to review it as well.

So we look at completeness but not necessarily responsiveness.

MS. WILBON: We don't look at appropriateness, I would say. We do look at whether or not they responded to the question, but in terms of the appropriateness and whether or not they put exactly or worded it in the way that we would like the committee to see, we generally leave that to the committee so that there's not a, well, why did you stop this? Because we've had the, we've heard that on the other side as well that we want to see what comes in. So I mean it's a balancing act then.

MS. DAMBERG: Yes. No, I understand that the committee wants to see it.

But I'm also wondering, you know, maybe there's sort of a middle ground here where there's a subset of the standing committee that does a quick review at the front end and sort of signals quickly back to the measure developers that it's not going to be sufficient to kind of make its way through the process in a seamless way. Because I think that there's a lot of time and energy spent here that maybe could have been sort of short cut at the front end.

DR. ASPLIN: Right. So I want to get feedback on the committee because I think this is a good discussion for the long run for how we approach this, and I'd also ask us to be parsimonious.

So Andy?

MR. RYAN: Okay, so just a couple quick points. Number one, I think the algorithm is quite good and reasonable and with respect to the measure we evaluated yesterday, you know, they did do reliability

testing.

There was a section in the appendix called reliability testing. They gave us some numbers that we can evaluate and we checked that. And with respect to validity testing, according to this we don't need validity testing for it to pass. It just needs to pass face validity.

So, you know, I think the measure yesterday didn't, even people didn't think it past that second hurdle, but I think what they provided was enough and it was responsive to what NQF was looking for.

I also want to make the point that with respect to the 20-page limit, I mean it seems irrelevant to me because the supplemental material can be how ever long the developers put in, and with the Yale application yesterday the section on reliability testing was one paragraph.

And, you know, maybe I would have liked to see more but that was enough. So

it's not like we need 100 pages. We just need a couple things that we're looking for. And so, you know, I don't think that's an undue burden for developers.

DR. ASPLIN: Thank you. Nancy?

MS. GARRETT: So I understand the attention here with the standard submission and wanting to have everything fit there. I found the visuals yesterday to be extremely helpful because of these complex measures, you know, that are measured over time.

And so I would just encourage you to think about if there's some way to build that into the standard form so that there's actually, if it makes sense there's a picture that you can actually look at of how it works. I thought that was really helpful to see.

DR. ASPLIN: Thank you. Jack?

MR. NEEDLEMAN: I really appreciate Taroon's frustration. And I share it a little bit, and I think it raises some questions about thinking through the process.

So we're here face to face. We've got this rule which says if we reject on one of these must-pass criteria we stop consideration. We get another round of voting. We've talked about what additional information we want.

I was not quite prepared to stop the discussion of this measure yet, which is why I gave the developers the benefit of the doubt on reliability in terms of testing. So that's one element here, to think about the process and how these votes influence that and therefore how it influences voting behavior. So that's one issue.

The second issue that I think is raised by this conversation is also do we believe the measure is reliable versus has the reliability been demonstrated? And we've seen enough of these other measures and I know how it's been constructed, and one of the reasons why these measures get to be not reliable is you've got outliers that sort of pull things

around and really change the rankings, but they eliminate that by capping the price per patient.

So I fundamentally believe this measure is going to be reliable in the sense of you do the split sample, you do the other stuff, you're going to get consistent results that would demonstrate reliability.

Have we seen all that yet? No, we haven't. But we've asked for more information that would provide that. Given that, given my gut feel that the measure probably is reliable, given what we've seen about reliability testing of similar kinds of data, I said let's get past the reliability and deal with the other issues on the measure, but I do that knowing that we've got another vote available to reconsider all this.

And that all entered into my decision to give the benefit of the doubt to the measure on reliability on this round of voting. But we need to think about how the

stop affects the decision to vote in a given way, and we need to think about this issue of demonstrating reliability versus believing the measure reliable at the committee level.

And the third thing is, with the development of these algorithms I'd like to see, you know, in some sense the algorithm incorporated into the guidance for the developers on here's how you're going to be tested, here's what you need to be providing.

So that's a third element in terms of looking down the road to future submissions and how these the algorithm can play in. And that was -

DR. ASPLIN: Thank you. That's very helpful. Gene, could you make your comment?

MR. NELSON: Hi, yes. Gene Nelson here. It's been a great discussion and a complex one. The suggestion was that in future that we ask the staff when they do their review and the TEP when they do their

review to actually use the algorithm for reliability and a similar algorithm for validity, and make their notes on the algorithm so that you can trace through the thinking and the conclusions of staff and TEP in advance of a discussion like this that we're having with people's, you know, various backgrounds and understandings about some of these technical issues.

And to treat it as a guideline would be treated at Intermountain Healthcare where the decisions are made but variances from the guideline are noted. and that allows improveability of the guideline for specificity and future use.

DR. ASPLIN: Thank you. Carolyn?

MS. PARE: I just wanted to go on record. I think I've mentioned to some of the staff that I think their support in this process this time around was particularly helpful.

In their comments, when we got the

staff comments though, they weren't necessarily giving us an assessment. They were clearly moving us in a direction for what to look for. I also found the TEP comments to be extremely helpful. I did, because I'm not an expert or a genius in this area I have a lot of paper that I look at and one is the very helpful document, that is, "What Does Good Look Like?" And we got that last year.

There was a lot of very helpful documentation that we all had access to and again in the calls could have raised questions where there was missing data because I think we knew that in advance. I don't know how you resolve that.

We all have only X amount of time to dedicate to this work, and I think that will always be a challenge. But I do, like I said, want to go on record in complimenting the staff and NQF for giving us the necessary information in advance of our meeting.

DR. ASPLIN: All right, thank you

for that conversation. So let's transition to validity and I'd first ask if either Andy or John had additional comments that I thought you had referenced most of the scientific acceptability comments. However, if you have additional comments on the validity section, Andy would welcome you to share them now.

MR. RYAN: My only comment would be that I'm not aware of any empirical validity testing that was done through this application. I didn't see any. And, you know, with the other application there was some formal process just that was face validity. I didn't see that in this application as well. Those are my only additional comments.

DR. ASPLIN: Thank you. John, do you have additional comments?

DR. RATLIFF: Yes, I don't have anything else to add from the comments. I think they've been covered.

DR. ASPLIN: Thank you. Bill,

from the TEP perspective do you have any additional focused comments on validity that you'd like to share?

DR. WEINTRAUB: Yes. So I'll comment from the TEP and my own thoughts as well. The TEP basically is an agreement with what Andy said that we really don't see much in the way of validity testing. So I think that the problem here with validity is if you don't know where you're going it's hard to get there.

So I have trouble even with face validity. When I looked at this and seen the data and I said, well, I don't know what that means. Is that good or is it bad? So I think that what is the goal? Is the goal here to reduce variation? Is it a goal to reduce resource use? If you want to reduce resource use when do you know when you get there, when have you gone too far? How much variation is acceptable? I think it's very hard to know.

Yesterday we suggested

opportunities for external validation. It's not clear to me that for a measure like this that there are the same opportunities for external validation. So this is not really a criticism of the developers. I think they're sort of trapped in the situation. It's not clear how with a measure like this you can know when you really have validity.

DR. ASPLIN: I think it's a good point. I would just add in direct follow-up to your comment that the market-specific conversations at least as they were constructed and took place with various plans in the Twin Cities market, and again going down to the next level from their plan level data with not standardized pricing but real information, it gave more context as far as where you stood locally.

It didn't answer the appropriateness question that's been raised in part of our discussions but helped you sort out both the resource use and then of course

you had indirect information about your pricing position that seemed to be at least a locally referenced valid approach to your relative standing. And I don't know if that's helpful or not, just reflecting on how the conversations go.

Jennifer, you have a comment on validity?

MS. HUFF: Actually my comment was pertaining to the last conversation, so I'll just hold off.

DR. ASPLIN: Okay, and unless there are not a lot of cards in the room -- I shouldn't have said that out loud. But why don't you go ahead and make the comment? I think that's okay. We'd like to hear from you.

Jennifer? Jennifer, you may be on mute.

MS. HUFF: Sorry about that. So are you saying it's okay for me to make the comment now even though it's not about

validity?

DR. ASPLIN: Yes, go ahead, please.

MS. HUFF: Okay. Sorry the timing with putting, when I had a comment didn't work. First of all, I just want to say I am really appreciative of all the work that both NQF has done and that the developers have done. I found the conversation very rich, deep and has really helped me, bring me closer to a better understanding of the measures.

I can say I think the process has improved significantly, so I think we're moving in right direction of getting to a better place of how to review these measures and assess them. I do think they'll always be a challenge because there is a lot of information to sift through and it's a lot of technical information. And that just is something that I think is inherent as a part of measure evaluation process.

For me, one of the things that I

was thinking about when I was evaluating these measures was our earlier discussion where we talked about how cost and resource use is still in its early development. It's more nascent than quality measures.

So considering that this is an evolution, and some of that played in my mind when I was reviewing the measures and I hadn't heard anybody else say that so I wanted to sort of make sure that was brought forward.

And then I'll just finally say I think the work of having the developers have a conversation with the committee before we met in person was really helpful. And I think no matter how hard you try to get the perfect form and try to have everything on the form, conversations really help and they really help in understanding.

So maybe there's more up-front work that still needs to be done before we get together in person so we're less surprised by some of the directions the committee is going.

Thank you.

DR. ASPLIN: Thank you, Jennifer.

Jack?

DR. NAESSENS: Since no one else seems to want to be jumping in on the validity issue, I appreciate the TEP's comments and some of the challenges of dealing with this measure.

So let me kind of step back and talk about how I think about validity which relates a little bit to the usability. And to me, when I'm thinking about the validity of these measures separate and apart from all the specific testing, there are three or four key considerations and concerns that I have. And having sat on these committees for awhile, those concerns are somewhat tempered.

One is, how complete is the measure of resources that are relevant to the illnesses, the diseases, the patients that are being reflected? And I always feel the need to say billed services are not necessarily the

most accurate measure of resources that are being used, but that's what we have in these measures consistently so I live with that.

But it's important to recognize what we're missing when we only look at billed services in terms of understanding what resources are being provided to deliver care. And so are the resources as they're being reported complete?

And the NCQA tells us repeatedly that, and we see the list of things that are being measured. They are complete. They've got drugs in there. They've got the behavioral health services in there.

So to the extent that we're talking about billed services we've got a reasonably complete set of billed services here, and that is sort of one of the first things that I think about when I'm, is this measuring resources? Well, within the limits of billed services it's measuring resources.

The second issue is the pricing

issue, and here we're using standardized pricing. And that has pluses and minuses, the limitations of standardized pricing in terms of understanding what resources individual health plans or individual medical groups or individual physicians actually have to organize and deliver care differ from what the standardized prices are.

Places that are heavily Medicaid that may have lower actual revenues per patient than places that are privately insured are going to have different resources even though the standardized pricing makes it look like those resource differences are smaller. That again is an inherent limitation of the measure.

And in thinking about standardized pricing I recognize that limitation but it hasn't been a bar to approving measures. It's just one of those limitations that I need to recognize and take into account when I'm thinking about what we've measured and what we

haven't.

So the standardized pricing methodology feels acceptable to me in terms of validity with all those limitations. The third is what we're trying to do here is differentiate between variations in resource use, service use that are not driven by the patient characteristics but are rather driven by differences in provider care practices.

So the third consideration is whether the risk adjustment or the model adequately differentiates between the patients that may need high levels of resources versus low, and that turns directly to the risk adjustment model.

And as I've looked the HC model for doing risk adjustment, it seems to me that based upon the report to CMS and some of the data that we've seen it is doing an adequate job of differentiating patients that need more resources from less.

So in general I find measures that

have used that methodology acceptable and have been okay with it. The way I would distinguish this measure from the ones we discussed yesterday is the HC method as its been presented and discussed in the documentation seems to do a better job of differentiating patients that need different levels of service than the way we saw the risk adjustment yesterday do that.

So again to me this measure rises to the level of adequate risk adjustment and differentiation of patients. The fourth issue is the interpretability of, you know, how do we interpret high, how do we interpret low? That to me falls into the usability issue and not the validity issue.

And we've been struggling with how to think about how to interpret resource use measures and recognized all the way from the beginning of this process with NQF that at some point we're going to have to link them to quality measures to get some sense of value,

but we can't do that if we don't have the building blocks.

So I've treated the incompleteness of these measures in terms of interpretability of how much is being spent as an inherent limitation at this point in the process, but I still want the measures for building block.

So I tend to discount that problem when I'm evaluating the validity of the measure. It's measuring something, how to interpret what it's measuring is a usability issue not a validity issue for me.

So that's how I have approached, to me, the key criteria here and why I find this measure meets the threshold of validity. The risk adjustment seems to be good enough. The scope of the services that are being priced are appropriate.

The standardized pricing, while I'm not always thrilled with it, I know how to interpret that and I understand the limitations of it in terms of thinking about

what's measured and what's not.

DR. ASPLIN: Thank you, Jack. Very helpful. Lisa?

DR. LATTS: First of all, it occurs to me that when we were doing disclosure of interest I probably should have disclosed that I was on the CPM way, way back when, when this was originally approved by NCQA. So just to get that out there.

Second, it is so much easier to be on these committees that are reviewing the condition-specific measures, because this is just so much harder. And it's different, and I really think that probably these aren't as helpful here as there.

That said, I agree with Jennifer's comment and Jack's comment, not only what he said just now which is far smarter than I could ever have something to say, but his previous comment about this being a work in progress and a building block especially.

DR. ASPLIN: Taroon?

MR. AMIN: I wanted to make a note about the introductory comments that Andy made. I just wanted to note that the requirement for validity, particularly face validity, is that it's systematically assessed.

So the face validity, it's not necessarily that we're looking to this committee to make a judgment about face validity and say, you know, it looks right or up or down, it's that the developer is submitting information that demonstrates that they've done that on their end and that's systemically assessed.

I just wanted to kind of point that out and make sure that we're sort of using that bar in terms of the face validity requirement.

DR. ASPLIN: Thank you, Taroon.

Lina?

MS. WALKER: This is a question for the developer. I was just referring to

your submission on what are the statistical results from the validity testing. You had a series of questions which you answered. And my read on that was that pretty much the results that you got were reasonable, that there wasn't anything kind of out of whack.

And so that was confirmation that this was a valid measure. I'd just like to give you the opportunity to say more about that, if there are other interpretations we should be gathering those sets of questions and results.

MR. REHM: I'll let Ben follow up, but just as a -- you know, it's interesting, and may expand your question a little bit. We built this measure for use in the real world, and I can't tell you how many times it failed because we didn't have the ingredients right, we didn't have the mix right, we didn't have approach right.

I was on the CPM as liaison at the time this measure first came out, and every

year we'd say, you know, close but no cigar. And then the next time it was, oh, well, that's interesting. That's a new wrinkle.

And I remember because one of the ways we were able to evaluate the measure each year as it came back one more time, one more time, was the stability. Because the early ones, it was really unstable. Plans were moving all over the place, and that became kind of our metric, if you will.

And so I think one of the, and maybe it's a problem we have. This is a measure we implement in the true space. When you read the submission form, in many ways really what you're reading is our story of implementation.

And the hard work of doing that and getting it right and getting it so that employers, plans purchasers, and, to some extent, because we do visual displays of the quality and resource use so that people could see grids, high-low, you know, things like

that for the consumer, you know, information.

So when we look at the data and bring in this annual report to the seat committee on Performance Measurement that reviews it every year that's the story of telling them we've tweaked this, this is what we saw this year. We modified this, we've changed the number of entity, number of members that need to be in the measure looking at standard error and we have that data in there.

And so I guess with a measure this complex over such a period of time kind of its arc of life, I don't know whether to call it an adolescent or a, you know, an unruly teen, but it's certainly getting closer.

And I think the feedback we've received from you, and we received maybe parts of this feedback in the earlier round in 2012, I mean I think it's been very, very valuable. You, just like our users, just like the people who respond to the public comment that I

referenced on three different occasions, are giving us information that helps us rethink.

I think in many ways we've been a little entrapped by implementation in thinking about that so much. We did not create the measure for NQF. I think you appreciate that. The user of the measure is not NQF. NQF has a terrific service to the quality environment and that's why we're here.

And that's why we're going to come back and we're going to try to come back with more information that's more helpful and does get at kind of like, what does good look like for a relative resource use measure?

Is that what the NQF provided us? No. It provides us, what does good look like on a kind of generic measure. And this is just a particularly difficult thing sometimes to translate. Sometimes we wonder if it's too hard to translate. It sounds like we're getting better at it and we need to improve and we've heard that message.

So I guess in answer to your question, yes, I think that there's so much looking at the results of this measure over time, but the lens that we look at it through is for, to speak to Brent's point, the usability and how it performs in the marketplace and is it telling a better story so that we can leverage it for other things. And the things we want to leverage this obviously for is to get at value.

Ben, did you want to add anything?

MR. HAMLIN: No, I don't have anything else to add.

MS. WALKER: Just to be clear and understand your response. So you were saying that you submitted this information as part of your validity testing, and so what you're saying is that the answers to those questions you asked kind of met the smell test.

So it was reasonable, in line with expectations of how plans should be performing on those various dimensions.

MR. REHM: Yes, I guess I was being so AHRQ-oriented, what I was trying to say is that different points in time that the story of its early failures really tells the story of its current status and that it did pass that test.

And each time we'd take it back to our committee on Performance Management and our Evaluation Measurement advisory panel, and they're listed in your submission form, they've said good work, keep it up, don't stop, keep improving it. And what we see here does not make us nervous or concerned about the validity of the measure.

DR. ASPLIN: We have two online and then Bill. Joe? Joe, did you have a comment?

MR. STEPHANSKY: I'm sorry. Who did you, do you want Bill first or me first?

DR. ASPLIN: Joe, go ahead.

MR. STEPHANSKY: Okay. I realized today that I have been around Jack long enough

to have been corrupted by him after serving on a couple of committees with him. I really like the way he laid out his four points, and I very much agree with him on that.

I look at the measure in the same way as him and think that the way he expressed that really belongs in our committee report as an example of how we have to deal with these kind of messy measures and where our limitations are.

Second, and I expressed this in an email earlier last night to some of the NQF staff. There's a Kaizen process that NQF went through, and one of the things that stood out to me in that was the necessity of measure developers telling more of the story. And that's actually what we were starting to hear from you in your last comments, for example.

And I just want to emphasize again to the NQF staff that that story is important to me and I think to some of the other members in terms of our final evaluation of the

measures and that we need to find a different way to get that story to the committee members. Thank you.

DR. ASPLIN: Thank you, Joe.

Gene?

MR. NELSON: Yes, two comments. One is, I think building on what was just said that the comments from the spokesperson for the TEP and then from Jack both indicated that for an expert in measurement to weigh in on validity and reliability there needs to be some sense of an operational definition specific to the case of cost or resource use. And that the specifications that Jack gave is an example, I think, of contextualizing what validity means in the context of this kind of measure. And it's very helpful.

And again going back to the algorithm, if the algorithm could have, reflecting the kind of operational definitions that are context-specific it might be helpful.

And then the second comment is

that towards the purpose of measuring value, costs in relationship to quality, point in time and over time, it would helpful if NCQA in its use could actually provide information.

And over the past four years, for example, how many plans made a substantial decrease in costs, where their costs were higher, and had quality hold equal or improve? Because value improves if a cost in this case were to decrease if you start at a higher position and if quality improves or stays the same.

So actually getting experience from the field in the plans on moving the parts of the value equation, what's the experience been going to the point of usability.

DR. ASPLIN: Thank you, Gene.

Bill, you might get the last word here before we go through the algorithm.

DR. WEINTRAUB: Okay. Well, I was going to sum up, so I think I might do just

that. I want to reflect on the comments I made and the comments Jack made because they're not in conflict, actually.

Jack talked mostly about construct validity and I think he's right. What I was talking about, sort of the overview of what does this mean, and that at the end of the day is a rub. But you've got to have something that makes sense as you build it up and Jack's right about that.

So Joe's comment and Gene's both related to construct a good measure, and then if you're not sure what all this means then look at it over time and see what's happening and tell that story, which is what we're beginning to get from the developer.

And when you put all that together this is probably as good as we can realistically get with this right now.

DR. ASPLIN: Thank you. Andy?

MR. RYAN: Just a quick point. So Jack took us through his criteria for face

validity which I think a lot of us think are quite reasonable. I think the question and the issue is that it seems like the developer should have gotten people like Jack in a room and asked the questions that Jack posed and then gotten their responses and then said people agreed with those things.

And then we would say that was a systematic assessment of face validity, and then we could say, okay, look, they did this and everyone thinks it's valid so we sign off on that. But, you know, absent that our committee is kind of making these judgments.

And so I think we're an expert committee and we are, I think, qualified to make this assessment, but it seems like NQF is calling for an additional level of testing to have been presented by the developer prior to that. So I think that's kind of maybe what some of us are struggling with.

DR. ASPLIN: All right, thank you.

So let's move through the

algorithm and the guidance from NQF for evaluating validity, and Taroon, I'd invite you to walk us through.

MR. AMIN: Okay, I will attempt to do that. I think one of the challenges I have just as a note, the conversation that we're having here is not the same tenor of the information that was presented in the documents by the TEP and by the preliminary evaluations.

So I'm going to try to summarize. I think I'll just walk us through this to say that, you know, number one, looking to see that the specifications are consistent with the evidence in support of the measure, I think generally the committee's okay with that.

I think we're all, I think now I don't have a clear sense of where the committee is based on the conversation and the information that was submitted in the preliminary evaluations at this point, so I'll

look for further clarification or not. You could just decide to vote.

But the potential threats to validity were empirically assessed, empirically assessed for exclusions, the risk adjustment, and then those are probably the two biggest ones. And the ability to statistically significant meanings to difference in performance may be less of an issue, but the first two appear to still be in. I don't see any question.

So depending on how you feel about those things that if you feel no that would go to insufficient. If you felt that that was addressed that would be yes.

And now we're getting to empirical validity testing, and again I think what I'm hearing from the -- again it's very difficult to assess this because there's differences of opinion, I believe.

But what I heard from the lead discussants is that there's some degree of

face validity, now the question would be whether it's the opinion of the group about whether it's systematically assessed. And again that is a requirement.

Andy's characterization of what's required in terms of NQF endorsement is a systematic assessment of face validity. This group isn't assessing the validity, the face validity of the work. It should be systematically assessed by the developer.

I don't know if that's sufficient, but --

DR. ASPLIN: Bill?

DR. WEINTRAUB: Well, I'm just wondering if the algorithm here isn't working very well, were all potential threats to validity that are relevant to the measure empirically assessed, the answer is no, then we'd have to write it as insufficient.

But it's too much to ask along the lines of our previous discussion. So I'm not sure that the algorithm's really helping us

here adequately.

MR. AMIN: So Bill, I think that's, I mean I think we're cutting to the heart of where there's a challenge here.

And I think we need to make a decision about this measure and maybe we'll move on, but we're coming back to that conversation after lunch. Because if we don't agree with the criteria then let's have that conversation and let's identify which criteria are either too high of a bar or not relevant to cost and resource use measures. Because we need to implement that consistently, and that's all I'm asking.

I'm not trying to say that this measure should go up or down, I don't have an interest or I don't, particularly, you know, the committee can make that decision. But my only interest in this equation is that we're consistent and that we're sending a clear signal. Now if we don't -- and so I'll just leave it there, and I want to come back to

that conversation though. I'll just say that, for the sake of discussion.

DR. ASPLIN: Lina?

MS. WALKER: This is a question for Andy just to get a sense of how the other committees viewed this. I did not see any systematic assessment of face validity, and I don't know if I'm missing anything.

But Taroon summarized what you said is that there was some disagreement, so was there some systematic assessment of this validity that I'm missing in here?

MR. RYAN: I can say I didn't see any assessment of face validity, and I didn't, I mean there was some committee member comments saying, you know, we think the measure is valid, but I don't recall seeing a comment saying, yes, the systematic assessment of face validity was sufficient or was done or whatever.

MR. REHM: And I'm sure we want to bring this to a close here pretty quickly. So

we're culpable in one area. This is a measure that is so, despite the seeming absence of data, you know, I can hold up this is our total data package, if you will.

And not that pages mean anything, but the due diligence is there. In a measure that was so complex and so difficult to develop in many ways reflects its complexity. We normally put in a two-page thorough ad nauseum description of our validity, face validity process, as Taroon comments, and it is systematic.

I've tried to touch on that during some of my comments, but if you don't mind and you have all the members of the eMeasure advisory panel, I think we may have failed to include our committee on Performance Measurement.

Most of you folks know many members on that committee, but we're glad to share that with you. Sorry it did not get into the submission form. I think we were so

taken with all the data that we kind of forgot that face validity might be kind of where we needed to be.

But just to do a short version, every measure that NCQA develops goes through a measurement advisory panel. You can take a look at the list. It's in this main submission form right at the end.

And I think you can appreciate this is the group that literally developed the measure along with staff. This isn't staff going and developing a measure then coming back and saying, oh, will you approve this. No, this was much intertwined because many of the people on that group have special skills, special talent and special motivation, if you will, to try to get a measure like this into the field.

So every year that I told you we brought this back to the committee on Performance Measurement which votes on the recommendations, and it takes two votes. One

vote to get the measure into public comment which goes to a 30-day public comment period, and again to vote it for inclusion, in this case an AHRQ/HEDIS fine for the year.

And then the first year measure's in play. It's actually in hold status. We don't publicly report that data. And then we bring it back for first-year analysis. And then that goes to the Measurement Advisory Panel again and then that goes to the committee on Performance Measurement and then that's the cycle.

This measure, because it's changed when we changed the risk adjustment approach to HCC so it went the first time around then it went for the HCC change, and then last year we took it because we lowered the number of people permitted in the measure, so to lower it and looking at the standard error, maintaining the same.

And then we also looked at the exclusions where we eliminated two of the

exclusions that are tested and are included in our materials. So in terms of that I wanted to make sure you were aware of that.

So this is, I think, then three cycles where it's been through the panel, public comment, committee on Performance Measurement. So in some ways I would say, and with votes by that and then of course votes by our board of directors.

So that's the full governance of NCQA. That's how we operate. That's where this measure went through just as every other measure we've presented to you has gone through that has the HEDIS imprint.

So I apologize that we did not include that boilerplate, if you will, but that's the process and I'm happy to answer any questions you might have about that.

DR. ASPLIN: Any questions from committee members about that process? Jack? Or other comments?

MR. NEEDLEMAN: Yes, so the face

validity in terms of the testing, I mean the committee members you've got are really good people.

But the two things that sort of, the three things that drive the results, you know, clearly the exclusions, and I'm reasonably comfortable with those and didn't recall the TEP complaining too much about the exclusions, per se. But it's going to be the decision to cap the maximum amount per patient so that pulls the variation in a lot.

And the second is the risk adjustment because the risk adjustment is where we distinguish between the variations due to practice and the variations due to patients.

So can you just speak a little bit to what kinds of analysis and what your conclusions were as you looked at the decision to use the HHC methodology, and then your experience using it in terms of how well you think it's doing right now in distinguishing

between variations in practice and variations in patients, and what the consequences are of capping at $100,000 and where that number came from?

MR. HAMLIN: So I can address some of those certainly. So, you know, as Bob mentioned every component of this measure had to go through a multi-committee process and that included the development.

The measures were initially developed in 2005. We didn't get to public reporting status until 2009, which meant there was about 14 rounds of development that went through this multi-faceted review before it was even deemed to be valid enough to go through HEDIS and the public reporting.

The initial cost caps were developed again using our research database where we looked at, you know, if you will, sort of faux calculations of the RRU in different scenarios and, you know, running different bootstrapping analysis in different

scenarios to determine what the appropriate cutoff would be based on the submissions that were there.

And I believe we revalidated that as still a appropriate cutoff in 2011 when we were also, you know, again we'd updated the database and refreshed it and then done some additional analyses for the exclusions for the eligible population size.

So every time we make what I would call major change that we're reducing eligible population size, removing exclusions, either we do a fairly thorough analysis and we basically, we take the entire measure back to these different committees to look at the change in the entire context of the measurement approach.

And so again when it was pointed out to us that, you know, these specific exclusions are very relevant to the condition, we looked at the effect on removing those exclusions across the entire measure, across

a number of different plans to see if there were unintended consequences elsewhere, and then those entire results were sort of represented back through the efficiency measurement panel, the CPM, the board, public comment, et cetera, et cetera.

So it's not just a matter of taking the individual adjustment to the board. It's kind of like we make an adjustment and then we take the entire measurement approach back up to these committees.

With regard to the HCCs, we actually initiated a testing of four different risk adjustment approaches to the RRU to kind of see which was the most appropriate or the most relevant to this type of model.

Two dropped off very early, and then so basically what we ended up doing a more thorough analysis on was the initial approach which is sort of a age, gender, comorbid, yes or no to the HCC, and it was found that the HCC was much more sensitive and

much more specific to this population using this at a plan level and it reduced the error ratios down to a level, you know, again to where we felt we could reduce the eligible population size. It didn't change those very much.

And again this was then taken back through this entire process in the context of the entire measure, what is this going to do, how is this going to affect populations, what's the effect on reporting.

The HCC was even bigger because it actually increased the amount of data required from the plans rather significantly because they were reporting in multiple cohorts. And so that was actually performed over a two-year period through multiple reviews.

And so that's sort of the way we approach each of these. It's not just a matter of tackling a change, it's a matter of, you know, the change in context of the entire measurement approach.

And we do retest and we sort of revalidate, if you will, taking it through all these different processes to make sure that aren't overreaching, unintended consequences that are resulting from that change.

DR. ASPLIN: Bob, did you have another comment? Okay.

So I'm comfortable with us moving ahead with the vote here. I think using the algorithm, the question before the committee really is whether the materials plus the subsequent discussion today would move us to a point of being comfortable with the measure having a systematic approach to face validity beyond empiric testing. So let's find out where we stand.

Evan?

MR. WILLIAMSON: We'll now vote on subcriteria 2b, validity. You have four options, high, moderate, low or insufficient. You may begin voting now.

And we have all the votes. We

have zero high. We have 17 moderate. We have one low and we have five insufficient. The measure passes validity.

DR. ASPLIN: Thank you. We're going to move ahead to feasibility, and again I would ask, beginning with Andy and then John, if they have any additional comments based on the committee's preliminary recommendations.

MR. RYAN: I don't have any additional comments. I guess I would say that as specified, you know, this is designed to be at the plan level so there's drugs that are in there.

You know, there were some, so for instance this kind of measure might not make sense for the Medicare population but that's not really the intention. I think that the overall comments were that the measure is feasible.

DR. ASPLIN: Thank you, Andy.

John?

DR. RATLIFF: The comments that were posted seemed to attest that for the plan level data, which is where the measure directed, it seems feasible and the data appears to be available and feasibility appears good, again at a plan level, that caveat offered.

DR. ASPLIN: Very good. Any additional comments from committee members or questions for the developers? Seeing none, let's move ahead with a vote on feasibility. Evan?

MR. WILLIAMSON: We will now vote on Criteria 3 feasibility. You have four options, high, moderate, low or insufficient. Begin voting now.

And we have all the votes. And we have 20 high and three moderate.

DR. ASPLIN: Thank you. Let's move on to usability. And again I would first turn to Andy and John to see if they have any additional comments on usability and use.

MR. RYAN: I think I would just, the overall, I think that people liked how the data were presented in the sample. The sample score sheet, that that was important for plans, the purchasers. It's helpful to have this information and there's a role for this.

I think there were some comments again about this measure, this level of analysis and whether and how actionable it was for the health delivery system, but if we're just kind of taking that as given this is a plan level, then I would say that it was a pretty widespread idea that this had high usability and potential use.

DR. ASPLIN: Thank you. John, anything to add?

DR. RATLIFF: I would agree with those comments. Comments from the committee seemed to focus on this being usable at a health plan level. I think if questions arise on terms it would be applied to a facility or individual physician level. I personally

would like to see more data or more testing with regard to that.

But nonetheless, comments of the committee were favorable with regards to usability of the measure.

DR. ASPLIN: Thank you, John.

Ariel, we're going to scroll back to a comment that we said we'd come back to. In your written comment earlier was can someone be more specific about what a health plan does with the measure? What do they find of value if no one can say high is good or bad? That was your written comment, and maybe you can expand upon that if you choose right now.

MR. BAYEWITZ: Yes. I mean it just seems like, you know, from the comments that people have been saying that no one has affirmed that directionally we know what do with the number, right?

I mean we are saying it's clear that it's saying something, I think, but we're

not sure exactly what that something is, right, so which gets back to Jack's comment. So I just want to understand. Because one of the earlier comments was, well, health plans have found it to be of value.

If someone could just walk me through sort of end to end, you know, the plan gets the number, the data that they see, again if the number directionally doesn't say something specific, what are they doing with it? How are they finding it meaningful?

DR. ASPLIN: Thank you. Ben?

MR. HAMLIN: Well, you know, again I think the measure results have a broader application than just the plans. So, you know, we've offered guidance on applications to identify cost opportunities to improve the numbers again, but we don't actually make any kind of recommendations that high is generally bad, especially for subservice category levels.

So, you know, we hear oftentimes

through the experience of going to these NQF committees about specific members who've used the measure structure for specific use and we've had a very informative.

We have not systematically addressed or, you know, systematically tested specific best practices or pilots that people have undergone based on their results from this measure.

MR. SAUNDERS: Ben, if I can jump in to add. But I think what we do know or we think that the measure provides tools to be able to assess the reality of the spending at the specific plan.

So we have in the whole suite of measures we're looking at that total medical spending for cardiovascular and we also have it broken out by the specific component categories, whether it's for inpatient facility charges or for the inpatient or outpatient components of E&M or procedure in surgery. We now have the lab in imaging.

So we have a broad spectrum of the components of cost, and so we feel like that by providing that information in the subcategories that we've set up an infrastructure for the health plans to be able to look at how their mix of services, what their observed spending is, and granted it's standardized but they're able to impute their own pricing to know what they've actually spent.

But they're in the position to have both pieces of information and they're the ultimate arbiters of the usability of this. But they have the component information to be able to say this is how much is expected of my spending for my population given how everybody else, all the other health plans across the nation that are submitting this measure are spending for similar populations. So we feel like the risk adjustment model provides a benchmark of sorts that is specific to each individual plan.

And so while on an individual measure, an individual metric, it may not be clear whether you should be up or down. In some of our papers we've found that greater spending on having a higher observed-to-expected is associated with higher quality performance on cardiovascular care and for diabetes care.

Think, well, gee, shouldn't we be encouraging people to spend less? But the benchmark there is perhaps as a mix of services that is being spent that we're underspending on a particular component where HEDIS is sort of a multidimensional service system that's contributing to the quality.

And so we think that by the plans looking at the components of services, looking at their own paired quality measures which are for the exact same defined eligible population that they're able to make those determinations for themselves of what actions to take either in terms of quality improvement or in terms of

how they choose to treat these populations.

So it's do we, you know, I think at sort of a basic level is you could spend less by doing bariatric surgeries or those types of things or you could spend less through exercise. That we see a variety of patterns of utilization that are consistent with high quality and we think that the health plans are in the position to evaluate in terms of their contracting and who they work with to make those decisions.

DR. ASPLIN: Thank you. Cheryl, then Andrea, then Tom.

MS. DAMBERG: I was looking at your documentation, in particular the sample report as well as what was in the table about the planned use for regulatory and accreditation programs, and I'm hoping you can comment a bit more on that particular application.

But when I look at the sample report you have that quadrant graphic where

you're pairing -- and utilization. And is that the type of feedback that you're providing on this, and as you move towards use for accreditation purposes, you know, is there going to be some signaling that if they do poorly on this they're going to receive a lower accreditation rating, or how does that play out?

MR. REHM: If I can just -- Ben, follow up on this. But so the graphic represents, you know, low cost/low quality, high cost/high quality, low cost/high quality, all variations on a theme.

And the intent for that was really to help both the consumer and the employer market, purchaser market, be able to understand, first, the variation around the cost and resource use, resource use and quality, to give it essentially an image that they could react to.

We held an employer forum around measurement a couple of years ago with many

top 200 Fortune plans, I mean companies, and I've got to tell you we were two days there talking about measurement. The RRU measure was the one measure that in our readmission measure at the plan level that just really caught their attention because it was trying to do this thing.

So in terms of, I mentioned before, the ability to take this into the accreditation program, which it's not currently in, is going to really be dependent on whether we can prove the point which you had asked previously which is can you demonstrate a true differentiation here, because that's what we require in order to benchmark and essentially rank plans on that dimension.

So that's the goal, if you will. Are we there yet? Not completely.

DR. ASPLIN: All right, so I'd like to take the last two comments here and then push through the vote so we don't miss our posted public comment time on our public

agenda, because that was at 12:15. So believe it was Andrea and then Tom.

DR. GELZER: Thank you. Just from a plan perspective, dependent upon the population, the plan population, I mean obviously it's going to vary dependent upon how much cardiovascular disease you have in the population. But that said, from a transparency perspective this is a valuable and usable measure to have in the armamentarium.

DR. TSANG: There's about 12 or 13 states that have legislated the use of all payer claims databases right now, so I'm just wondering whether this measure will be somehow connected to those efforts. Because, I mean that process also by the states are doing, not doing this, but they are doing similar comparisons between Plan A-Plan B.

So I just want to understand that usability of this measure in the context of what the states are doing and if there's

redundancy or there's any parallel efforts.

MR. REHM: I can't speak to our policy department. I have a hard enough time doing measure development. But, you know, we certainly observe that first thing that the all payer claims database holds a lot of promise. Absolutely they do.

In the context of a cost and resource use measure, many of the states, depending on which ones they are, have distinctive limitations on the use of that data. I think in our own minds we would love to have real costs, you know, imputed into this so that it's more proximal to HealthPartners Total Cost of Care measure and, yet, conveying the quality dimension as well.

So I think that it is a better thing to have all payer claims databases out there, and it would be great if they could loosen up the restrictions on some of the use of that data from a policy perspective. That's not a measure development thing. We

would be promoting that and advocating for that for the purposes of true transparency.

DR. ASPLIN: Thank you. Let's move forward with our consideration of usability and use. You have the options in front of you. We'll move to the voting, and Evan, let us know when you are ready.

MR. WILLIAMSON: We'll now vote on usability and use. You have four options, high, moderate, low or insufficient. Begin voting now.

Okay, we're still missing one vote. Is everybody still in the room? Yes.

And we have all the votes. We have eight high, 14 moderate, one low and zero insufficient. The measure passes usability and use.

DR. ASPLIN: Thank you. And we'll move to our final overall suitability for endorsement, yes/no. Does the measure meet NQF criteria for endorsement? Evan?

Excuse me, comments. Nancy?

MS. GARRETT: I think this was the section where I was supposed to bring up the stratification issue. I apologize. Can we talk about that now, or is this the wrong time?

DR. ASPLIN: You know, let's -- I don't think that's going to affect this vote. Let's do the vote, let's see if there are public comments, and then you can make that comment after that if that's okay with you.

MR. WILLIAMSON: We'll now vote on the overall suitability for endorsement. You have two options, yes and no. You can begin voting now.

(Off the record comments.)

MR. WILLIAMSON: So we're still waiting on one vote in the room. If everybody could please try one more time.

And we have all the votes. And we have 20 yes and two no. The measure passes.

DR. ASPLIN: Thank you. Nancy? Or excuse me. Are there any public comments?

Let's move to that first. Thank you.

MR. WILLIAMSON: Do we have any public comments in the room? Operator, could you please open the lines for public and member comment?

OPERATOR: Yes, sir. To make a comment please press star then the number 1. There are no public comments at this time.

MR. WILLIAMSON: Thank you.

DR. ASPLIN: Nancy?

MS. GARRETT: So my proposal is that I want to see if the committee would be interested in making a recommendation that this measure be stratified by sociodemographic characteristics.

So the developers presented evidence that there are associations between race, ethnicity and gender and utilization on this kind of general concept of heart disease care. And right now again this risk adjustment committee is making a recommendation in June and that will possibly

change the current policy that NQF has which is that those factors can't be used in actual risk adjustment.

But the current policy does allow for the committee to recommend stratification by those factors which means basically reporting by particular groups. So I wanted to get feedback on whether people think that's something we should comment on.

DR. ASPLIN: Committee comments? Jack?

MR. NEEDLEMAN: Yes. I think the policy context for thinking about this is critical. We're seeing a major expansion of Medicaid managed care. We're seeing major expansion of insurance with many people going into limited, you know, into HMOs or exclusive panel plans where issues of adequacy of the networks have been relevant and where adequacy of non-physician services in the community have been critical for thinking about the consequences for both health status and both

outcomes and use of other kinds of things like readmission.

We heard Andrea yesterday talk about Zip Code as being the critical determinant of whether you got readmitted and -- I'm sorry, was that Janis? Okay. Well, you're both on that side of the table.

So, you know, to the extent that the kind of data are supposed to serve a reporting purpose to understand the challenges facing different plans and also the public policy purpose to understand what the challenges for committed providers to deliver care in different communities or to different populations, I would encourage more analysis and more presentation of data that allows us to understand the SES factors associated with the ability to get needed care and get appropriate services.

DR. ASPLIN: Ben, related to Nancy and Jack's comments, could you clarify a comment you made earlier around from the

feasibility using existing data whether plans could stratify? And I thought you made a comment that some of them are not collecting the required data to do it systematically, or did I miss that?

MR. HAMLIN: No, that's correct. I mean that last assessment, which I think was two years ago, there was still far too much variability in the plan data for us to require reporting, where we're trying to make strides in that direction as Bob alluded to and we are certainly open to recommendations from this committee about, you know, future ways to present the results.

So we're certainly happy to look into it, and like I said we are currently waiting very patiently for the SES and sociodemographic factor recommendations to be coming out.

DR. ASPLIN: Bob?

MR. REHM: Nancy, we were just talking about it during the break. I think if

we were to venture forth in looking at these new data elements in forming a measure and its interpretation we would probably start simple.

I would not want to, I'm not sure I'd throw it into this particular measure first. I think there's a lot of learning curve on how to do this and do this well. And it could very well be we might start with some of the component quality measures that link to the cost and resource use just to get a start.

But I mean, this is a big lift, a big lift downstream, and we're all aware of it and we want to do it right. This will not be a tomorrow thing. It'll be maybe a few weeks after tomorrow.

MS. GARRETT: I mean I would just respond quickly. I mean this is the measure before us so we can't comment on your other measures right now, but I think that the conversation that's happening nationally and locally about this issue is really different than it was even a year ago, and I think that

we might want to consider actually making a statement about it.

Now what that really means if we were to make such a recommendation, you know, I don't know. But I think as a committee we certainly could choose to do that.

DR. ASPLIN: Yes, certainly, without really settling the question of appropriateness that's been raised during this morning's conversation, and also might begin to inform some of the questions about whether higher is better or lower is better, et cetera, and that dialogue.

Larry, you have a question or comment?

MR. BECKER: Yes, I do. This is Larry Becker. So I agree with all the comments that were just made, and I think, you know, not for this measure, but I do think that it's an important thing to begin to look at in maybe in terms of subsequent measures. Because it seems to me that we need to

approach care in a more patient-centric way to provide different, maybe it's an opportunity to provide different groups with approaches that we can do and that they're able to do.

And so maybe it provides us an opportunity to get some leverage into care that can actually be followed.

DR. ASPLIN: Very good. And Jack?

MR. NEEDLEMAN: One more thought on this issue. I talked about the policy context and the community context. But the other issue that occurs to me is you're using standardized pricing, and I understand why and I think there's a lot of value in seeing standardized rates.

But I also noted that in the discussions about standardized pricing in other settings we've raised the issue that it hides things including real differences in the resources that different plans have available depending up on who's contracting with them and at what rates.

So you're not, you know, for all kinds of proprietary reasons you say you're not getting that data from the plans and I understand that. But some kind of SES stratification either by the Medicaid versus others or other kinds of SES stratification may help us understand where implicitly some of the differences in resources exist and make more apples-to-apples comparisons of plans with comparable levels of resources available to them in terms of their performance, not only on the quality measures but on the resource use measures.

DR. ASPLIN: Janis?

DR. ORLOWSKI: Just a quick comment to tag onto what Jack's saying. I'm surprised that a first step wouldn't just be Zip Code data. And I'm sure the plans have that. And, you know, that would be an initial foray into taking a look at some stratification.

MR. REHM: You know, I'm familiar

with Zip Code for assigning race and ethnicity status and even language, and I know there's been a lot of work done by RAND and others. Cheryl probably could comment on it.

But, you know, Zip Codes are interesting. Zip Code home, Zip Code point of service, Zip Code hospital, you know, it does get what sounds so simple when you peel it back. But I mean, I think that as a starting point --

DR. ORLOWSKI: We're not talking about the hospital --

MR. REHM: Right.

DR. ORLOWSKI: -- being socioeconomic. We're talking about patients.

MR. REHM: Right.

DR. ORLOWSKI: And I think that what we're talking about is comparing services that are provided to the patients.

MR. REHM: Yes. No, I

DR. ASPLIN: I'd like to thank the committee for the conversation throughout the

morning, and a special thank you to the developers. Thank you, Ben, for joining us. Bob, Robert, for participating with us in this conversation.

I hope that some of the takeaways for it enable the similarly constructed relative resource measures in other clinical conditions hopefully will go even more smoothly. We'll see.

So with that we're going to break for lunch and reconvene at 10 after 1:00. We do have some give in the remaining elements of our agenda so we should be able to get out on time so everybody can get to their travel plans. So we'll reconvene at 10 after 1:00. Thank you.

(Whereupon, the foregoing matter went off the record at 12:42 p.m. and went back on the record at 1:18 p.m.)

A-F-T-E-R-N-O-O-N S-E-S-S-I-O-N

(1:18 p.m.)

MR. WILLIAMSON: So we have another afternoon discussion today, so we thank everybody for their attention and comments during the measure evaluation section. We really think that was a very rich discussion and I know we covered a lot of topics that we'll address in the report, and we'll definitely be coming back to you in future meetings, and there's a lot of information there.

So rather than kind of doing a deep dive on stuff this afternoon, we really want to kind of circle back to some of the efforts we've been making to improve our processes and get some feedback on that as well as maybe revisit a little of the discussion we had yesterday as far the role of the standing committee, how we can use this committee to kind of push things forward for cost measurement and for this area.

MR. AMIN: And then can I just jump in real quick? Oh, well, you're going to --

MR. WILLIAMSON: Yes. So most immediately we have Phase III, and I know that you think Phase II has just started but Phase III is starting up now too. The measure submission deadline for Phase III is April 18th, so these are staggered but they're kind of overlapping.

We have another orientation call scheduled for the 23rd. Again we're going to think about how we're going to use that as far as a standing committee goes to make sure we're not repeating information that we just went through, but really try to use it to make sure that we cover some of the issues that we have identified during this phase as far as measure evaluation and moving forward.

We'll do the same thing. We'll be convening a TEP, a pulmonary TEP, still thinking about how we're going to consider the

dental measure, but we definitely need to convene a pulmonary TEP to provide input to the full committee.

We've gotten a lot of great feedback from Bill about how the TEP went for the cardiovascular process, so we'll be considering that and making sure that we've got some good input to the committee from the Technical Expert Panel.

Of most import is the in-person meeting. That's June 25th and 26th. That's all been scheduled. The dates are posted on the SharePoint site. We'll be sending out the calendar invites for all the Phase III just to make sure it's all on your calendars.

I think all these dates have been sent out in some form or another over the past few months, but we really want to make sure that we, we can get on this early and make sure it's on everybody's calendar and everybody knows what's going on. So we'll make sure that these dates are on there.

So again this will be wrapped up by the start of 2015, so again this is a quick trip through cost and resource use measure and for these projects, and as far as future phases we don't have anything concrete yet but we're definitely going to utilize the expertise of this group going forward through your terms.

Are there any questions about Phase III? I want to make sure, you know, it's crept up on everybody so we'll make sure that we get it on everybody's calendar.

All right, so during lunch I distributed a survey, and I also sent out an email with a link to the survey. In case you don't want to fill it out on paper you can type in into a SurveyMonkey, but we will accept the paper survey. We've already gotten a few of those.

And really looking to, you know, we've made some big changes, some subtle changes and we're really looking for feedback

on that. We want to make sure that we take into account the committee perspective, and we also have an analogous survey that we've given to the measure developers.

We really want to get all perspectives on this, all the stakeholders, everybody we're bringing to the table to make sure that we get their feedback on how things are going.

And so among the items on the survey we have how we handle orientation, the workgroup, or for this committee it was Q&A calls as far as the measure documentation. I know that this phase was a little different than last time as far as the way you receive documents, the type of documents you received and how it was distributed through SharePoint.

We did a lot of work on redesigning those project pages, but again we really want your feedback. You guys are the ones, you ultimately have to use the material. We want to make sure that we're making it as

easy as possible for you guys to participate with this.

We really appreciate all the volunteer hours you guys put in and we want to make sure it's as easy as possible and that, you know, we're not wasting your time or doing anything that doesn't, you know, we're not introducing any waste, I guess, going back to some Lean principles.

So in that regard we want to go over the staff reviews. So what we definitely meant at this time that was different was providing some staff input on how we think the developer addressed the questions, and really just more identifying things to look for, not necessarily directing you any direction but making sure that there are, you know, we can really focus you in on certain key issues for the measure documentation.

And how we handled the TEP review, we really wanted to see were those questions appropriate? Did you feel that those

questions led you to a good answer? That it led the TEP to useful information for you as far as getting input on the clinical specifications that you might not necessarily be as familiar with?

And then how we handled the preliminary evaluations. You might have noticed this time rather than submit the high, moderate, low ratings before the meeting, we got rid of that. We just wanted general comments. We felt that the high, moderate, low that were submitted before necessarily didn't, they didn't correlate with what ended up happening at the meeting and we really didn't think it was that valuable of an exercise to go through that.

We just really wanted to make sure you guys started thinking about the measure and going through, so we really want feedback on that. That's something that was very important.

So I'm teeing us up right now. I want

to open it up for discussion. I just want to go over some of the slides and then we'll open it for anything that's on here as far as some verbal feedback.

Okay, and finally is the meeting facilitation. That's something we really put a lot of work into as well. I know you might have noticed that we designated two seats for the measure developers. We really think this was more of a conversation with the developer than in the past where they've kind of been, you know, off in a corner and only called upon at certain times. So we want to get feedback on that as well.

We've got a lot of feedback from developers as far as their interactions with the steering committee, so a lot of this as a result of that feedback through our Kaizen process and other feedback we received from developers, we're really trying to engage them. You know, make sure that this process is valuable for them and that they want to

continue to participate with NQF.

So I guess I covered a lot there. So I want to open it up for general comments. Again we definitely want the survey feedback as well, but if there are general comments right now about kind of the changes you've noticed.

We have some new members. If there are things that you all want to address with us while we're going through this, we definitely want to hear that now. So we'll go ahead and start. Nancy?

MS. GARRETT: Well, this is some minor comment, but I really did like having the measure developers kind of sitting at the table with us, talking to us. And it would be nice for them to have a name card because after they introduced themselves it was hard for me to remember who they were.

MR. WILLIAMSON: All right. I posted the names of the developers on our SharePoint site just so that in the future

we'll make sure that we do that just so everybody knows which of the developers.

That information is listed on the measure information form, but again is buried and sometimes you don't necessarily know if the person who filled out the form is the person who's presenting in person.

I think those are kind of a last-minute thing to make sure that those two seats were saved. I think we usually have placards, but again with our offices being closed on Monday we were kind of scrambling yesterday morning to get everything printed.

But that's definitely a great point. We will make sure that we do a better job of that. Cheryl?

MS. DAMBERG: First of all, I want to thank all the NQF staff for putting everything together. I know how much work this is to put together these packets, and I think you have been working really hard to make it easier on committee members and we

greatly appreciate that.

I want to second Nancy's comment about having the measure developers here. I think that was really important. I particularly liked the pre-call where we got to ask them questions.

My suggestion that I made to you at lunch time was possibly thinking about making that a mandatory call for the committee so that people can voice a lot of the issues in advance of us coming together. Because I think there was a lot of time spent here that maybe could have been dealt with earlier in that call to try to move things along faster.

MR. WILLIAMSON: Great. Thank you. Brent?

DR. ASPLIN: I want to compliment staff on the changes to the measurement packet, the measure packet that we received. I think it really helped clarify and prioritize the area to focus on and I found it helpful both, plus in the interaction between

the TEP and the staff comments, you know, I kind of, it really helped zero in on what I should be looking at and offering an opinion on.

And in the setting in particularly of large, complex measures, and maybe they're all going to be large, complex measures, I don't know, that was helpful to me. I think that's a significant step forward.

MR. WILLIAMSON: And actually I'll push that a little further. So one of the things that we did was try to make it an evolving document, and we understand that has some challenges as well. So I want to raise that where we started with the staff review, and then when we got the TEP feedback we added that to the document and then when we got the preliminary evaluations we added that to the document.

And we know we've gotten feedback that some people like to print out the documents and then they don't know which

version they've had, so I just wanted to get some feedback on the best way to be able to share new information as it's added. So I don't know if anybody has any thoughts on that.

MS. WALKER: I didn't have any problems with the way that you had shared the information. I think it's pretty clear. It's all dated on the SharePoint site so it's clear when you loaded it and it was clearly labeled so you know what the document is and you can get to it fairly easily.

You were also again, once again to join everybody else who's already said this, but I think staff has been outstanding. Really, the review for the three measures were very, very helpful. The questions that you posed, you weren't trying to influence our decisions but you were trying to provide a frame for us to think about these measures. I thought that was terrific.

I like that you constantly, and

you were great in constantly reminding us of various meetings. That's very helpful. I know everybody else is as busy as I am so it's nice to get those gentle reminders. I really appreciate it. I don't think it's overwhelming. Keep doing it. So I nothing but good things to say about the staff.

MR. WILLIAMSON: Great. Thank you very much. I know we have John on the phone and then we'll get to Janis here in the room.

DR. RATLIFF: Yes, just briefly. I was a bit reluctant to use the SharePoint initially, but you guys did a great job of organizing the files, keeping them updated, clearly showing the timeline of when the files were being generated and posted. And I found that portion of the process to be extremely helpful.

And again, I hate to echo the crowd, but I really commend the NQF staff for what a smooth and diligent job you've done with organizing the standing committee.

You're to be congratulated for it.

MR. WILLIAMSON: Great. Thank you. That's very helpful. Janis?

DR. ORLOWSKI: I would consider this a very minor comment because I found, being a new person and looking at the material for the first time, I thought it was highly organized. And the SharePoint worked fine and, you know, very, very organized for me to be able to figure out and use the guidebook that you gave us and stuff like that.

So I thought it was probably one of the first times that I've been on a committee where you really don't learn what the committee does eight meetings later. I mean you actually gave all the instructions to me.

So to my minor point, I was on the first call and there was a lot of discussion, a lot of stuff that was going on. And so I went back to the SharePoint site afterwards and looked at the transcript and I would say

it's nearly impossible to get any information because it is literally a transcript, and so you get all the ahs and ums and everyone going through.

And I was looking for two pieces of data regarding a conversation that I recalled and had trouble finding it. So what you have on here is both the transcript and the recording and I don't know that we need both of them. But what would be great would be, you know, sort of summary points.

And again that's work for the staff and I apologize for that but that's my, and as I said those are minor comments but the SharePoint was terrific.

MS. WILBON: So I have a question. This group has been a little bit different, I will say, from our clinical committees because they have significantly more measures, so 20 to 25 measures per committee generally for our clinical areas.

And the unique aspect about this

group is that we've had less measures but they tend to be a lot more complex and so, you know, we spend more time on each measure than our clinical committees would.

They have a process because they have more measures where they divide their committee up into workgroups and have a series of calls before the in-person meeting to really kind of figure out what those key issues are, and then the work of the workgroup, if you will, comes to the in-person meeting and so they spend less time.

Again they're reviewing 20 to 25 measures so it's much more of this is what the workgroup highlighted, and then the committee as a whole really spends time focusing on what the issues that the workgroup teased out.

We haven't used that process as much because we have less measures. It's a lot harder to figure out how we would divide people up into workgroups, you know, only having three measures. And so we've somewhat

repurposed those pre-calls, but just wanted to get your input on whether or not you think that workgroup process would be useful.

And seeing that we only have three measures, you know, the in-person meeting time might be, you know, would we need a whole two days to do an in-person meeting if we had those pre-workgroup calls since we are kind of working through those issues at the meeting as opposed to on the phone?

So I think these are some of the things we've been trying to balance out because this group is a little bit different. So just your thoughts on that will be useful and we can potentially implement that for the next, you know, phase of work that we have.

MR. NEEDLEMAN: I do think the work of this committee may be a little bit different, but it's also because we're in a very different stage in terms of the experience with measure development here.

So a lot of the discussion has

been around core issues of how the measures are done. We were talking at lunch about, you know, risk adjustment. There are three or four standard ways of doing it and a bunch of other ad hoc ways to do it.

As we all get more familiar with the standard risk adjustment methods and their strengths and limitations, the adequacy of risk adjustment conversation will go faster. On the other hand we probably want to think about ways to capture or memorialize some of the background discussion of things like risk adjustment, things like exclusion rules, so that new members of the committee will get a chance to learn from the experience so they can get up to speed faster, and we get a little bit more chance to reflect on what we've said about things in the past so we can be bringing a little bit more consistency to our evaluation of the measures.

So I'm not quite sure where that gets done or who or how that gets done, but we

ought to be thinking about some of the consistent issues that we have spent a half hour or 45 minutes discussing and how to capture some of the issues that have been raised and the points, the decisions, seem to turn on so people have a framework for looking at new measures using the knowledge about how we've evaluated things in the past.

MS. WILBON: We had a discussion with Janis earlier and it seems like you guys had a similar thing about the transfers and, you know, how do we handle transfers. Who gets credit? Which hospital gets credit for the transfers, and is there kind of a general principle that we, you know, as a committee so that next time we see a measure that uses exclusions or inclusions related to transfers, is there a way that in general the committee feels that that should be handled for resource use measures?

And I think it's something to think about related to the other issues you

were discussing as well, Jack, of whether or not we have a way for you guys to set up some principles or something about these kind of overarching issues and the way that we should be kind of framing our discussions around the measures that come forward so there's a consistent approach.

MR. NEEDLEMAN: Right. But the other thing we were also seeing is sometimes after discussion we've reached some consensus and sometimes after discussion we haven't. So the attribution rules on all the per member per month measures were contentious. They remain contentious. I'm not sure very many people's views on whether they were appropriate or not changed very much.

So those are the areas where we know it's going to be contentious but we don't have to, you know, we've had the discussions over and over again. So it's a matter of, you know, understanding how they've dealt with it here, so we need to think about -- but no, now

I've wandered all over my tongue here. Let me try this again.

In some cases we have real disagreements about what the appropriate standards are to apply, and the documentation of where we've been and what kinds of rules we've used should reflect that. And in other cases we've developed a little bit more consensus about these issues and we can probably, you know, revisit them periodically rather than routinely.

DR. ASPLIN: At risk of reopening the whole conversation that we had right after our vote on reliability, I guess -- let's reopen it. We didn't really want to go home today. No.

To the extent that it is going to be an iterative process between developer and committee, and to the extent that we would take Cheryl's recommendation that the call with the developer would be a key step prior to the in-person meeting, then I think we need

clarity on what levels of information we include in our voting. Because I've gotten, and I'm just trying to put it out on the table if you want to improve the process.

I still don't know if we're trying to stick to what's in black and white or if we're supposed to include additional information, and if so, what levels of information are deemed appropriate. And then at the end of the day how do we stay consistent in how we approach those questions?

Taroon? Yes, that would be great.

MR. AMIN: There was one thing that I wanted to add that I think is a good bridge from where we were and where you're going. And I just want to provide the committee with kind of a macro context of why we think that this process needs some improvement and what sort of the intention of this work is.

So it's generally, this is a very intentional process. So as many of you may or

may not know, the majority of these consensus development projects are supported through our federal colleagues who are, you know, these projects are generally run through contract and the timelines and the funding, the majority of it's supported by our federal colleagues with an understanding that we're able to bring together you as members and experts on these topics to come together.

However, the reality around the fiscal environment and the pressure to do things faster and more efficiently is certainly present. However, particularly in this area of measurement and I think those of you that have, I mean it's broadly an issue, but it's particularly acute in this area of measurement, cost and resource use and particularly in readmissions, the process of getting toward consensus is it takes time.

So one of the questions that we were working through as an organization is how do you make sure that you have good voice in

the process from the membership, the developers, the steering committees in particular, but also have a sense of efficiency in the process? So that's one macro objective.

The second macro objective is that we recognize that just an up or down decision on measures is not very satisfying for the committees and it's not very satisfying for developers, and most importantly it may not be getting us the rapid amount of change that we're going to see in the next few years in where we need measurement to be, and particularly it may not have the effect that we need it to have.

And so one of the principles that we've employed through our Lean Kaizen efforts is that, you know, we make some of these recommendations but these recommendations may be so far past the development of the measure.

So, you know, do we really expect that the measure developer is going to have a

fully-specced measure after as Bob mentioned, you know, spending a million dollars on development and then we're going to convene a committee that's going to say, you know, this part of it needs to change and expect that the whole cycle is going to start over again?

Realistically that's not the best optimal use of time, but there should be much more upstream guidance or upstream input from a multi-stakeholder group on, you know, what are the development priorities, where do we want to see measures, and have a much more long-range view of development in various different spaces.

What's really unique in this particular area of measurement is that we are very new in the sense of the number of measures that are in the cost and resource use domain, so it gives us the unique opportunity to start off on the right foot so that we're not sort of retroactively saying, well, how do we clean up the portfolio and are these really all the

measures that we need? And a huge amount of development dollars have gone into building an infrastructure for measurement and we're not really sure that it's actually resulted in any improvement.

So very thoughtfully we're asking the question about how do we build infrastructure up front to make sure that we are giving guidance to HHS around where they should be investing their development dollars in terms of gaps and priorities, and also sort of setting a strategic vision around the complex measurement issues that may arise so that we don't have all these dollars spent on the back end in terms of trying to implement components that can't be implemented, whether it's because of concern around attribution or the fact that, you know, that's an easy one to pick on, but you can imagine there's other sort of complex measurement issues that need to be addressed.

So as we're moving forward, the

important thing is to sort of think about, yes, we're doing the up and down on measures, still, and this goes directly to your point, Brent, which is, you know, well, how do we think about the information that's in front of us? Because ultimately the committee does still have responsibility to make an up or down decision.

And right now the answer is we don't really know because we're trying to transition our process to a place where we can be more iterative with an understanding that we're trying to effect a much more upstream process.

And so we're looking for guidance about how best to do that and we're also looking for developers to play a different role with us and be open to that type of relationship which we haven't really deliberately, you know, spent time building in the past.

DR. ASPLIN: If I could just, quick follow-up to that then. A

recommendation. I think there's power in using an iterative process. I think you get to a better place by going back and forth and using an iterative process.

So my recommendation would be to expand the list of potential data sources that committees will be making decisions on. I think the key will be how do you make that transparent and so that people who aren't involved in the process can understand what data were made available that were not in the original submission. How did the committee move down this path to get to where they got so that it doesn't seem like a black box and it doesn't seem random.

So I would say if we can figure out a way to capture the power of an iterative process while making it transparent, so that if there are questions about the process not just the decision those can be raised, I think that would be the sweet spot. I'm not saying it's easy to do.

MR. WONG: So Taroon, I appreciate your broader view and I could tell that staff took a lot of time developing both the guidance documents. And you can see that, you know, it's been very thoughtful about the whole process.

I think that one of the things I think that caught probably the membership here by surprise was, in fact, that document existed. So if it was in some of the books already then I totally missed that. And so when we were going through the process of looking at reliability and validity, you know, we're paying a little bit more attention to the guidelines there.

And I think that that's where part of the conversation kind of emerged about, well, if we do this what will happen to the measure. Because we don't want to quite, you have some concerns but you don't want the measures taken totally off the table.

And so as Jack kind of mentioned,

it was kind of a strategic loading situation especially for today given our experience yesterday. So that's a broad comment. And I totally understand the need for, as you put it, consistency of how we do things, so I think that we do need to be mindful about how to best deploy this particular document. So that's one point.

My second point is that for today's measurements, you know, you mentioned before and this is one thing I wanted to ask, but I think that staff kind of alluded to this that for the measure developers there is information there about what the threshold is. And for this particular measurement I kind of wonder whether or not they kind of missed something in terms of that, because clearly there was not enough information in terms of the reliability and validity. It seemed to be missing.

And, you know, I sit here thinking, well, they've been doing this for a

long time. It passed the first level. But we were just missing that other set of information and it was really hard for us to make that call on it because we know from experience they have been doing this. So there has to be something there but we didn't have that information.

And so I kind of wonder how much of it was really on them for not paying enough attention to that sort of thing in delivering all that information versus us trying to tease that out and try to come to that other place.

So it might be a process issue where, you know, going back to the developers, some comments may be even from staff that, you know, from sitting on a lot of these committees I'm pretty sure that this particular issue is going to merge, so stand ready.

And, you know, part of my comment here is kind of, if you think about the federal grant process and if you have a good

grant officer that is looking at applications coming back, you know, the grant officer goes to all those study section meetings and they're sitting in the background and they're listening to all the issues that kind of come up.

And in some way there's this little role that, you know, I've been through this a lot, you know, you might want to pay attention to these sort of things, right. And again it's up to either the applicant to kind of make that decision to take that advice or not. So just a very broad comment about how we can potentially fix that process.

MR. WILLIAMSON: Great. Thanks, Herb. Tom?

DR. TSANG: Two comments. I echo everyone's comments about the thoughtfulness of staff in preparing the guidance documents, so thank you very much. I guess this is more about kind of like the life cycle of the measure.

And, you know, Taroon, you really talked about the Kaizen process, and I'm just kind of invoking my conflict of interest here maybe. You know, in the pharma world we constantly do post-market surveillance and look at adverse events of drugs in the real world.

So I'm just wondering, you know, I'm sitting here listening to the measure developer saying, well, you know, this measure's gone through puberty and it's like reaching adolescence now, but it would be nice to actually think about either the unintended consequences of this measure and have a little bit of that data, I think some of us would actually benefit from that.

And if he can present the almost like a growth chart of this adolescent and tell us, you know, where has the measure pivoted in terms of its specifications and also, you know, what were the consequences or the impact whether positive or negative, so

that I could learn whether, you know, it's been useful the last three years. This has gone through a re-endorsement process and I think that type of information would be helpful.

And then the whole process about looking at adverse events is whether a measure developer would want to take that, you know, I know resources are constrained and funding's lacking, but to look at the post-endorsement process a year from now and collect that data about the measure.

So it's about really trying to refine the measure process and refine the measure development process as well as the improvement process of a measure. So I'm looking at this from a standpoint of like drug development as well.

So, and then I guess, you know, this is totally aspirational in terms of the technology platform to capture this data. I know SharePoint's a little bit clunky, but at

some point, I guess, if you have the resources to create a technology platform where you can actually combine all the different data points into the worksheet.

MR. WILLIAMSON: That's a great point. So my list, right now I have Lisa, Bill, Nancy, Cheryl, Andrea, and then Larry. So we'll go with Lisa.

DR. LATTS: Thanks. So some of our meeting over the past couple days reminds me of those old AQA meetings we used to have about seven or eight years ago, and some of you were there with me. And we used to have our knock-down, drag-out, yelling fights about how to evaluate these clinical quality measures.

And those just don't exist anymore because we've come so far in measure development on the clinical side that we know how to evaluate clinical measures. I don't think we know how to evaluate these measures yet, and I don't think we can apply the same

criteria in what we've learned on the clinical side to these measures.

So I think the criteria need to be different, and I don't know what that is exactly. I don't know if it's just that we loosen some things up, and I especially don't know how to do it and still provide the consistency that we're looking for. So that's what I'm struggling with.

But I think we need to somehow recognize that these measures are different and we don't have the same level of, that we can't do the same level of testing and it's not nearly as straightforward, even though I know that it's still not straightforward on the clinical side, it's even less straightforward on this side. Number one.

Number two, I thought, Tom, your comments were excellent, and I wonder if again especially for these measures there needs to be a special set of questions that are asked for the recert measures. So it doesn't just

go through the normal process, but there's, you know, maybe there's some things we take off.

You know, we don't do the importance to measure, but maybe we just do is this, still, you know, is this measure still relevant? And then we do, you know, what's happened since you implemented this? Who's using it? How are they using it? What are the results? What are the problems?

And so we remove some of the stuff that's clear in a recert measure because it was approved the first time and we add some stuff that allows for some of that in-depth evaluation.

MR. WILLIAMSON: That's great. Thanks a lot, Lisa.

So Larry, I know that your point was to that. I don't want to go out of order, but do you want to add anything about the different criteria? I think it's something we'll have to bring back and go in-depth at a

later point, but Larry, do you want to add anything to that?

MR. BECKER: Yes, I just wanted to lay the idea on the table that for, you know, most of the history of NQF we've used these endorsement criteria around clinical measures.

And as we pivot towards, you know, cost and resource asking the question of whether or not these are still, all of these are still the appropriate criteria and are there other criteria we should be thinking about as we pivot to these kinds of measures.

MR. WILLIAMSON: Okay. That's great feedback. We'll definitely have more work on this. We'll definitely talk with you guys about this. So we'll move on to Bill.

DR. WEINTRAUB: So I'll reflect on the last several comments. While the measures we use in resource may turn out to be somewhat different, there are still principles of looking at reliability and validity that will pertain and we will have to do those things.

And I think we're going to have to help the developers develop a tool kit to be able to respond looking at reliability and validity.

For the measure we heard about this morning, the comments all started off with, well, they didn't look at reliability or validity. And they really didn't. Jack put it together for them. He put together their face validity and their construct validity for them.

Would have helped a lot if they had done that, and they could have if we had a good framework for them. But there are other criteria for validity to think about, the criteria of validity and consequential validity, and there's a formalism to it that people should go to.

They should respond to each of those even if very briefly, saying, you know, we don't know what the consequences of this are going to be, even if that's all they've got. But at least let's help them develop a

framework that then would help us. I think we could do a better job.

MR. WILLIAMSON: Thanks Bill. We'll go to Nancy.

MS. GARRETT: So also building on some of the other comments, and Taroon in particular, your comment that in some ways we're giving input kind of too late, I wonder if the standing committee provides an opportunity to look at this a little bit differently and do something a little more creative around an iterative process.

So it would be really nice if all of us really became experts on cost and resource use measures knowing what's out there in the community, what's being used, what's being developed, what's in the pipeline. The things that we've looked at in the committee, how are they being used.

I mean, could there be a monthly newsletter with some information so that we understand what's going on and could there be

some more regular check-ins, rather than once every three years this really deep dive and then we don't look at it again?

And it's the same thing that came up in the episode grouper criteria discussion, which is if you endorse a software product that needs to change every week what does that even mean? And so we don't want these measures to become static. So is there a way we can somehow change the process to be more iterative in response to that?

MR. WILLIAMSON: I think that's a great point. Next, we have Cheryl.

MS. DAMBERG: So I agree with everything that's been said so far, but I want to turn to kind of another issue. So at the beginning of the meeting you had presented a slide that looked like the boxes within the box to go from resource use to, what was it, efficiency to value or something like that?

And then you had another table that showed us a set of measures but it didn't

include all of the overuse measures that you have evaluated on the quality side. And I think it's hard for this group to kind of keep track of all the moving pieces that hit this space and sort of what boxes we have actually filled versus where are these gaps such that when a new measure comes forward we can say, oh, you know, this is a resource use measure for hospitals and, you know, that's a space that we don't have anything in.

So I think just being able to get our heads around that in a more systematic way would help the committee think more clearly about, you know, what is it that we're trying to accomplish here and where is this going.

And I think that this has been the challenge that the MAP has faced because, you know, we want to be doing this at all levels of the system and we want it to cover these six dimensions and, you know, it gets hard to think about.

But I was really struck by, I

think, Jack's repeated use of the word building blocks. And so if partly what we're reviewing here is a building block to even get us out of that first box, I think that's something we have to acknowledge and say is that sort of the basis on which we're reviewing this measure? That we're really in the alpha stage of development rather than beta into scalability.

MR. WILLIAMSON: Taroon, did you want to respond directly to that or should we go to Andrea?

DR. GELZER: I guess a question and a comment. And the question is, how many cost and resource use measures -- the portfolio's very small. What is the pipeline and how are we soliciting folks to develop these measures?

MS. WILBON: That's what we were asking you. No, we're -- and Taroon talked about this, I think, a little on the first day, but right now our process has been very

reactive in that --

DR. GELZER: So how can we become proactive, I guess, is what -

MS. WILBON: Right. Working with you guys to figure out what are the gaps, what are the high priority areas that we should, to help us, you know, work with developers and some of the people that we know are giving the development dollars to tell them this is what you should be spending your money on. Because otherwise our work will continue to be reactive to what's already out there as opposed to -

DR. GELZER: Well, so maybe you should formally solicit, or formally send that out to the committee so that we can each tell you what we think.

MR. AMIN: Right. So maybe I'll just piggyback on where Ashlie was going which is that these are sort of the enhanced functions of the standing committee that we see. We're sort of putting them out for

reaction, but we're certainly not trying to answer all these questions now.

So one of them is in its domain which is, you know, what are the high impact areas of cost and resource use measures that we would want to see by care setting or how would we start to look at the clinical areas, how do we start to address these?

We won't be able to get -- how do we start to prioritize the clinical areas where we want episode-based measures? How do we start to think about episode first per capita measures and which cases would one be more appropriate than the other?

And so we're just sort of putting these out there as questions that need to be answered, and we need to think through on our back end, which is how do we start to create this as part of our work going forward?

And part of our conversation now is to create an expectation with this group that this is where we see this group moving

to. Not just up or down on measures but, you know, we've been doing some of this work.

I shouldn't say, you know, definitely this group has not been just up or down, certainly, but again because it's been such a new area and there's been a lot of conceptual building, conceptual model building that we've done from the beginning.

But we're going to have to go through that exercise, Andrea, in a much more systematic way.

DR. GELZER: Okay. That's great. And then my comment, really, I think that we can't -- the way the process is set up now it's probably too prescriptive for this area. I agree with what Lisa said that this is a new area. We have to consider these measures a little differently than the quality metrics and then evolve and iterate how we do the evaluation.

And I would just say in a past life when health plans were first starting to

talk about efficiency measures and I had to go out on the circuit and go to specialty societies and talk about our plans, you know, new high efficiency network and how we were measuring physicians on efficiency, we have to make sure that we are able to use the same vernacular or the same terminology as folks out in the community. Because all I remember is, well, tell me what's the confidence interval? What's your confidence interval?

I mean, I think, you know, we need to be speaking the same language that they will be speaking. And that's just an example.

MR. WILLIAMSON: Thanks, Andrea. So I have Janis, Bill and then Matthew.

DR. ORLOWSKI: So what I would say is that I view this as a spectrum of healthcare from physicians to hospitals to, you know, the ambulatory to the healthcare plans. And I think what we have to do is, if you're saying where shall we go, I think the answer is that you have to take a look at that

continuum and be strategic where you believe, based on the experience of the group that we have, where you believe it would be critical for evaluation of cost and resource.

And I'll give a example. Large hospitals have a lot of experience looking at cost and resource use in order to improve efficiency. And what they have learned over time is that sometimes you go to, you know, you go immediately this is a high cost item or whatever.

And then as you look at the continuum what you understand is that there are times that you spend a certain amount of money and it improves everything downstream. It improves your costs. It improves your resource utilization, your length of stay, and it improves your outcome.

And so what I would say is that if you want to say where you can make an important contribution, take that example from within the hospital and use it on, you know,

as sort of a broader look at healthcare.

And I think that there are areas that can be pinpointed to question cost and resource use and it can be coupled with quality. I will also say that there are areas along the spectrum of healthcare resource usage where there's clearly an underspend by folks.

And I think that it would be very interesting to not only take a look at areas where we believe that there's high cost and high resource utilization, but it would be very interesting to take a look at places where there's likely low resource use, inappropriately low resource use. And I think that you can look at the spectrum and you can begin to target that.

And so, but that is a strategic discussion over, you know, a couple of days where you take a look and then say where is it likely, where are you likely to be able to make a difference?

And I think that there are people both on this committee and others that you could add that would have a lot of experience in looking at and evaluating strategically resource utilization.

DR. WEINTRAUB: So I'm going to build on what Janis just very wisely talked about. We're just barely beginning to sort of scratch the surface here and we're doing it without a framework to plug that into.

So the three measures we discussed these two days to me didn't fit into any kind of framework, they were just individual measures. And to develop that framework, that strategic plan to do that and to think about how we're going to serve society best in looking at cost and resource use we're going to have to have a framework. Because there's so many different things that you can look at. We've sort of, in a sort of standard way looking at providers and episode of care in the hospital in 30 days, or a health plan

looking at cardiovascular resource use.

But Janis just brought up the whole issue of healthcare disparities and how shall we address the issue of cost and resource use when socioeconomic factors and healthcare disparities are so tremendously important in our society?

Where's the balance between acute care focused on largely an older population and preventive care for children and young mothers? How do we as a society come to a balance in looking at resource use and that? How can the measures we would develop here help address those kinds of issues in our society?

So I think it's just going to take time and we're going to have to have time where we step back from individual measures to consider just what we're doing and how we can most efficiently use the limited resources we actually have here.

MR. WILLIAMSON: Thank you.

Matthew?

MR. MCHUGH: So very much, I think, in this same vein, I think in order for us to identify gaps we need to have a landscape to really look at it, and it just can't be we have this measure, that measure and this measure.

It needs some kind of organizing structure in order to say, okay, well, this is what we're covering here and how we're covering it, you know, but there's this whole other kind of black hole that we're not really approaching.

So whether it's a framework or starting out, I think, as Cheryl kind of talked about in at least mapping things on to that kind of very general kind of orientation of building blocks towards value would be a good place to start and would help us get the most value out of our collective thinking.

MR. WILLIAMSON: All right, I have Tom, Brent and then Dolores.

DR. TSANG: Two more comments. I was pretty disappointed in the measure developer's response to my question about the states' APCD initiatives because his response was we're only measure developers, I don't touch policy.

Well, I think that's the wrong attitude to have because obviously this is a huge impact on policy and it's inextricably linked to policy. So when he's telling me he's spending a million dollars on this measure development and yet this whole huge initiative is going on, so I'd like to understand why they haven't really, you know, coordinated activities.

So I don't know if that's within your scope, but I think that's one issue. And then the second issue is really coming from NQF's conference a couple weeks ago about patient-centeredness.

And so far, you know, this is the second committee I've sat on, but so far these

measures are either payer-facing or provider-facing, and we've been giving lip service about measures that would be patient-facing.

And so some of you may know about Castlight. It's a new company that's, it's not so new anymore, but I think they're thinking about going on IPO and they're presenting quality data along with cost data to all these employees about plans on the state exchange.

So I mean, how is that, you know, how are they doing it in such a way that could actually synergize and coordinate on these types of measures that are looking at cost and present it in such a way that's going to be consumer friendly?

As I think we all know that high deductible plans are becoming the majorities on these benefit plans now these days and that data is important, that information is important to consumers, so how is this measure going to be kind of like patient and consumer

friendly?

MR. WILLIAMSON: That's great. Thank you.

MALE PARTICIPANT: Great point.

MR. WILLIAMSON: Brent?

DR. ASPLIN: Yes, the conversation is sort of drifting naturally to the other broad question I was going to ask, which is what are the, you know, practical and best guesses as to what the next steps might be to the conversation we had just before lunch yesterday?

And it really kind of echoes in several of the recent comments around the need for a strategy in this area. And I know there are funding questions, yet who's being held accountable?

How will we actually hold medical groups to accountability or some combination of medical groups and hospitals, not just hospitals or plans? The two-tailed question of not just over, but to Janis's point, under

use of resources, and some of the other questions that have been raised that really lend themselves to more of a strategy process.

So will that, you know, and I know you might not have the final answer today, but what's the likelihood and who are the potential sources to fund that type of effort? It almost seems like a collaborative effort between a standing committee and the MAP, potentially.

MR. AMIN: Can I just add to that list real quick? Because I think that just as we talk about strategic issues, I think this criteria issue is on that list too. And then related is this whole issue of the fact that there's no directionality.

It's not clear whether higher or lower is better and how do we, you know, and that obviously has a link to the whole quality aspect. So those are at least, I agree with that strategic list.

MS. YANAGIHARA: Yes, that was

actually my point. This is actually a combined point for Cheryl and I, but Cheryl had to leave. So, you know, she pointed out that with clinical quality, I mean all the measure development really was focused on clinical evidence, and where was their clear evidence and let's make measures around that.

But there's not that with the resource use measure. Which is better? We had that conversation many times over the last two days, right, like what number is the right number? Is higher better, is lower better? It might depend on who you are so as to answer that question.

So what I actually think is that in a way resource use makes the most sense when it's paired with quality, and so if you look at the key quality areas, and I think yesterday that list that was up there was pretty good and it was really kind of played out in our own data those were the top areas.

It's really like finding resource

use measures that can then be paired with the quality measures. And I think someone else brought up there are already some kind of resource use measures that are considered quality measures like C-section rate, right? That is considered a quality measure, but really is also a resource use measure. It's not cost but it's resources.

So anyway I think that if you focus on those areas that clinically are important, then there's context for understanding the resource use and what is kind of better or worse because you have it to pair with the quality.

MR. WILLIAMSON: Thanks a lot, Dolores. We have Jack next.

MR. NEEDLEMAN: Not going to be my most clearly formed thoughts, and which given some of the thoughts I've had it's a very low bar. I was struck by Janis' comment about, in essence, value stream mapping which is what I took from what you were saying.

We've got usability as our criterion and we've going over it rather quickly. But in terms of setting priorities for what we need to look at, what's useful to look at, both the resource use measures and the quality measures are fairly high level endpoints of what in terms of summaries of how well the organizations are doing or the individual clinicians are doing.

But to make sense of them and to think about feasibility we need to think about we're trying to tie both of them back to what happens in clinical practice and administrative practice.

What's the value stream map that allows a hospital to efficiently produce care? What are the linkages to the post-acute services that makes sure those services are delivered appropriately to the patients that are coming out of the hospital?

You know, if we think about it in terms of treatment things like cardiac, we've

got the issue of what's the care map? What's the care map look like for an optimal treatment of this patient given the uncertainties, you know, and the probabilistic things that happen to patients?

And the usability of these measures are a way of testing whether you're on those paths at a very high level and whether they give you signals about where in that path you should be looking to improve yourself.

So at some point I think, and doing that linkage I don't think is in the NQF space, but in terms of setting priorities and thinking about the value of these measures, we need to spend a little bit more time thinking about the usability and the uses and how the measures relate to the uses.

And then the reliability and the validity come into do they give you accurate information? But so those who are making use of the measures should be part of the

conversation about what the priorities are, and what the shape of the measures are, and what kinds of information gets distributed, along with the summary statistics from the measures in terms of helping shape what a good measure is and what a good set of analyses are.

MR. AMIN: So Jack, one of the questions that we're thinking through internally is, does the use case for the measure change the criteria or does it just change the way you would potentially weight the criteria?

MR. NEEDLEMAN: There are two issues here. One is you've asked us where the priorities should be. How can we be more proactive? And I think the areas of priority and proactivity depend upon use. Who needs a measure to help them make improvements in care? So that goes beyond the evaluation of the individual measure.

The second question is, you know,

we've looked at some of these reports, that we saw it with respect to the NCQA report. We saw it in some of the stuff that CMS was thinking about distributing. We saw it with the CMS per member per month measure that was evaluated earlier, where the report becomes part of the usability assessment but also provides some insight into what the measure has to accomplish.

So maybe we ought to be reweighting the evaluation of the measures between, are they valid and reliable which is where the core activity, and if we allocated our time the last time few days where we spent most of our time, and the usability measure. What's the value of this measure in use to clinicians, to administrators, to plan administrators in terms of enabling them to make improvements in care?

As we get more measures and we come to that issue of, you know, the relative value of different measures which is down the

road, I don't think we've hit it here yet, but it's one of the criteria in the NQF list, I think this issue of relative usability and value to user is going to rise higher in our decision making and maybe we can figure out ways to anticipate that and send some signals of that to the measure developers as something that they also need to be spending time worrying about rather than simply is it reliable or valid.

MR. WILLIAMSON: Thanks Jack. So we have Nancy, Lisa and then Janis.

MS. GARRETT: So two comments. One is on the -- Tom reminded me of something when he brought up Castlight which is, I wonder if that's a stakeholder group that we don't have represented that we should which is analytic vendors who are really doing some very creative things in this space and might be able to inform some of our conversation. So that's just something to think about.

I know on the episode grouper

committee we do have quite a bit of representation from that kind of stakeholder group. But it's almost like an emerging group that's doing a lot of innovation in the measure of costs and resource use measurements. So it's just something to think about.

And then on Brent's question of what would it look like to get the resources to put together a strategic plan for cost and resource use measurement, I'm just wondering the NQF staff reaction to that.

So do we need additional resources? Could we do that without an in-person meeting, for example, through kind of working together virtually? What are thoughts on next steps for that?

DR. LATTS: So I just wanted to add to this schema that we're building of sort of the things we would like to see as part of this committee, and I know I've reiterated it multiple times.

But I think depending on the measure and the type of measure that you're looking at, I don't think cost and quality are enough. I think we need that measure of appropriateness or a measure of utilization as that third leg of the stool.

And so, you know, again for global measures it's less important, but when you start to parse it out into particular pieces I think it does get to be really important. And, you know, I know that Daniel Wilson, who was on the first, or the last phase of this committee, whatever phase that was, and who leads the Choosing Wisely initiative, always says that this is not about measurement it's about professionalism.

But I think those are, you know, his list is a good place for us to start. And we've got to start having measures of appropriateness in terms of what's being ordered.

MR. WILLIAMSON: Thanks Lisa.

We'll go to Janis.

DR. ORLOWSKI: So, Jack, I feel like you were reading my mind, and so just to continue my discussion. So I want to use, hopefully this is helpful and not starts a whole other part of the conversation, but when I take a look at the measures that we have before us in the last two days and we take a look at total cardiovascular spend, I see issues with that. You know, like anyone I could be critical. I could see it.

But as far as usability, I think that it has some, I think that it can be directly usable to plans, to hospitals, to large physician groups. I see a lot of applicable use for that.

When I take a look at heart failure, even though I agree that it's the number one diagnosis in hospitals and it is absolutely a critical issue that we have to address, the pressure point for heart failure is not the acute hospitalization.

The pressure point for heart failure is the ambulatory care of a patient with heart failure. And so it's right horse, wrong rider is what I would say. Should we talking about heart failure? Absolutely. Should we be looking at measures on heart failure? Probably one of the most critical measures that we should be dealing with.

If you take a look, and again we could be wrong, if you take a look at the pressure point for making changes in expenditure in heart failure it is not the acute hospital admission and 30 days post. It just isn't.

And so at some point for those of us in the field, you've got to, you know, you bring 25 years of experience to an understanding of how to utilize this data. And in my opinion that's how I look at these two measures. One is a little fuzzy but usable, the other is more discrete but is at the wrong usability point.

And so I think that Jack, you're right. That we're talking about value mapping and we're talking about clinical care mapping and then where do those become usable and where do we have the ability to influence our use of resources wisely.

MR. WILLIAMSON: Thanks, Janis. So we're going to Lisa, Lina, Dolores, and then we're --

DR. LATTS: I just wanted to push back on that a little bit. CMS paid for Yale to do that measure because CMS is the one paying for the inpatient hospitalizations. So what's their point of leverage? So I 100 percent agree with you that it's change in the primary -- or the cardiologist's office, whoever's caring for it, it's an outpatient change that needs to occur.

But who's going to make that change happen? All of the stuff that's happened up until now has not made that change occur in the outpatient arena, whereas holding

hospitals accountable will cause them to push it downstream so that that change happens.

DR. ORLOWSKI: CMS pays doctors.

DR. LATTS: They don't pay them enough. And it's the leverage.

DR. ORLOWSKI: Well, that's another issue.

DR. LATTS: But no, no, no. But it's the leverage. They pay them so little, frankly, that they don't have the leverage.

DR. ORLOWSKI: So Lisa, I completely disagree with you. Putting the pressure on the hospital to be the deep pockets to affect change in the ambulatory space or the provider space or the plan space is interesting. It works to a certain point.

But it's, you know, it's not where the pressure should be put. CHF is an ambulatory sensitive resource use. Now if you say hospitals are beginning to own more doctors then, yes, you know. But again it's a physician-specific point that is sensitive

here.

MR. AMIN: There's a lot of disagreement on this topic. Let's just maybe move. Maybe we could just move around. I mean, I'm just trying to be respectful of people's time to be able to get out, just so that we can get through the list, Evan.

MR. WILLIAMSON: Okay, we'll take one or two last comments on it and then we can, or we'll give Lina the last word on this and then -

MS. WALKER: Well, I'm not actually going to speak specifically to the heart failure issue, but I think Janis raises a broader point that I wanted to mention which I mentioned yesterday.

I think part of the issue with the heart failure measure was because the developer was trying to align it with the previously available quality measure. And so sort of backward engineering and measure so that it would complement something they

already have, I think, is a wrong approach to building a cost and resource use measure.

So once again I'd like to, you know, we really need to step back and ask what is the problem we're trying to solve for and think about what is the quality utilization measure and what is the complementary quality measure needed to address that problem?

And I actually do agree with your point about heart failure but I'm not going to go into that. But I think it highlights that problem that we're facing now is that they base it on whatever they have on the quality side and that's really not the right approach.

MR. WILLIAMSON: Yes, so we'll get Dolores and then Andrea, and then we'll call it.

MS. YANAGIHARA: So Taroon, you asked about whether the criteria should be different based on use case or whether just the weighting. I think the criteria have to be the criteria. I don't think that those

should change.

But the way that they're applied in terms of how much you value each thing or what is the rationale for a certain exclusion might be different based on different use cases. But I think the criteria need to be consistent, I mean just from a -- yes.

And then just a note on Choosing Wisely measures, Lisa, I agree. Actually they're not measures yet, that's the problem. There's a lot of great concepts but there's no measures yet on choosing wisely. So I think it's great. That would be a great area to focus on.

My concern is that they might be too narrow to get really robust measurement because usually they're very specific to a very small population of people. But I think it's still worth pursuing, because I think those are really important areas that getting to that appropriateness area.

And then just a note. When you

get into utilization measures and cost measures, it's tricky because if it's not paired with value-based purchasing what doing the right thing may be meaning losing money.

So for example, reducing C-section rates might be the best thing for the mother and the baby and it might be the best thing for overall total cost, but the hospital and the doctor are going to lose money on that, right?

So until incentives are aligned, it's hard. There's just conflicting signals that we're providing in the market. And so it's just something that just makes it, you know, the resource use measures that much more tricky.

MR. WILLIAMSON: Thanks, Dolores. All right, Andrea.

DR. GELZER: Okay, one more point. We had the discussion at dinner a little bit about who the driver is, and I agree with Lisa that the hospital in this day and age is the

driver.

And there's way too much fragmentation in the system and so costs may be, you know, what we're trying to accomplish here is rationalize the cost equation and to rationalize that we've got to reduce variation, but we've also got to reduce fragmentation.

So I think that, you know, for congestive heart failure, by god, the hospital is still the biggest cost center and they are still the biggest driver and we've got to use that. Thank you.

MR. WILLIAMSON: All right, thanks, Andrea. So we don't want to stunt the discussion. We invite you to send us emails or maybe we could find out some way to use the SharePoint. That we could implement a discussion board on there so maybe we could start seeding some ideas on there and really get some good discussion. You guys have a lot of opinions and aren't afraid to share them.

So we want to make sure that we give you guys a forum to do so and we can really start moving forward.

So the last thing we want to go over, just the next steps, we'll do public and member comment right before we close, but here are some next steps for Phase II and Phase III. So we plan to have a draft report posted by April 21st. Our post-comment call will be June 4th, so that's the next time we'll all be together is June 4th on that call for Phase II.

For Phase III we have an orientation call on April 23rd. Again as I mentioned earlier, we'll figure out really a high leverage way to use that. We don't want to just reiterate the same information that we have before.

We have Q&A calls scheduled May 28th and June 11th, and again, you know, maybe some of the feedback we got today might influence how we use those, whether it's a

workgroup call or true Q&A call. Question?

MS. WALKER: Now this phase, the first Q&A call happened before the technical expert panel convened. Is it possible to have both calls after?

MR. WILLIAMSON: Actually the first Q&A call happened after the TEP met. There were things we had to stagger that moved, you know, we were starting all these processes. We had the first TEP call and then we had the first Q&A call, then we actually had a second TEP call scheduled that we cancelled.

But in order to turn around the information from the TEP we didn't have it available in time for the first Q&A call.

MS. WALKER: Oh, I see.

MR. WILLIAMSON: So we can do our best to make sure that it's available, but again we're going to rethink how to use these. These are the times we have scheduled. They're on the books. And we'll think about

how we can use them to make sure that we get the TEP information to you in a timely manner. But again this is all, these are compressed timelines.

MS. YANAGIHARA: Do you have the times for these meetings yet? Because I'm like, I don't have them on my calendar or -

MR. WILLIAMSON: Yes, we'll send out -- the Phase III calendar invites will go out. They're also all listed on the SharePoint page. I'll just show you where they are here. We have the committee calendar. And so all the times are listed here, so I think they're all noon Eastern. All those calls are noon Eastern.

And see, they're separated by phase here. But we'll make sure we send out the calendar invites, but we want to make sure we get it on the calendar. And then our in-person meeting is June 25th and 26th.

So those are the next steps. Again we realize this is a compressed timeline

and we're kind of starting Phase III while Phase II is still going on, but that's just how the timeline lays out. So are there any final questions or feedback on that? We want to give you guys some time to get to the airport or get on your train.

So we'll close with public comment. Do we have any comments in the room? Okay, do we have -- operator, could you please open the lines for public and member comment.

OPERATOR: If you'd like to make a comment please press star and the number 1. At this time there are no comments.

MR. WILLIAMSON: Great. Thank you. Well, we really appreciate everybody coming to Washington and braving the weather and working with us over the last two days. We really, I think we got a lot of work done. A lot of questions to answer going forward, but we know we're all up to the task. So thanks again, and I want to thank our co-chairs.

DR. BURSTIN: And also thanks to those on the phone. I mean, for most of us who can't handle more than two hours on a conference call, the idea that they've hung with us for two days is really above and beyond the call. So thank you.

DR. ASPLIN: That was the comment I was going to make. They did a much better job than I could have possibly done hanging in there on the phone. So I really appreciate that. And one last thank you to the staff here for coordinating everything. Thank you.

(Whereupon, the foregoing matter went off the record at 2:35 p.m.)

H

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 Phase II

Before: National Quality Forum

Date: Wednesday, March 5, 2014

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