NATIONAL QUALITY FORUM
+ + + + +
COST AND RESOURCE USE
STEERING COMMITTEE
+ + + + +
THURSDAY
MAY 9, 2013

+ + + + +

The Steering Committee met at the National Quality Forum, 9th Floor Conference Room, 1030 15th Street, N.W., Washington, D.C., at 8:30 a.m., Eugene Nelson and David Penson, Co-Chairs, presiding.

PRESENT:
EUGENE NELSON, DSc, MPH, (Co-Chair), Dartmouth Institute for Health Policy and Clinical Practice
DAVID PENSON, MD, MPH, (Co-Chair), Vanderbilt University
BRENT ASPLIN, MD, MPH, Fairview Health Services
LAWRENCE BECKER, Xerox Corporation
MARY ANN CLARK, MHA, Interalign
CHERYL DAMBERG, PhD, 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
DAVID GIFFORD, MD, MPH, American Health Care Association
LISA LATTS, MD, MSPH, MBA, FACP, LML Health Solutions, LLC
MATTHEW McHUGH, PhD, JD, MPH, RN, CRNP, FAAN, University of Pennsylvania
MARTIN MARCINIAK, MPP, PhD, GlaxoSmithKline

JAMES NAESSENS, ScD, MPH, Mayo Clinic
JACK NEEDLEMAN, PhD, UCLA Fielding School of Public Health
CAROLYN PARE, Minnesota Health Action Group
DAVID REDFEARN, PhD, WellPoint
ANDREW RYAN, PhD, Weill Cornell Medical College
JOSEPH STEPHANSKY, PhD, Michigan Health & Hospital Association
THOMAS TSANG, MD, FACP, Harvard Medical School
LINA WALKER, PhD, AARP - Public Policy Institute
WILLIAM WEINTRAUB, MD, FACC, Christiana Care Health System
DANIEL WOLFSON, MHSA, ABIM Foundation
HERBERT WONG, PhD, Agency for Healthcare Research and Quality
DOLORES YANAGIHARA, MPH, Integrated Healthcare Association

NQF STAFF:
GERRY SHEA, Interim President and CEO
HELEN BURSTIN, Senior Vice President, Performance Measures
ANN HAMMERSMITH, General Counsel
TAROON AMIN, Senior Director, Performance Measures

KAREN PACE, Senior Director, Performance Measures
LINDSEY TIGHE, Project Manager, Performance Measures
ASHLIE WILBON, Senior Project Manager, Performance Measures
EVAN WILLIAMSON, Project Analyst, Performance

Measures
CARLOS ALZOLA, NQF Statistical Consultant

ALSO PRESENT:

RICHARD BANKOWITZ, MD, MBA, Premier Healthcare Alliance

JEFFREY BALLOU, Mathematica Policy Research

CHAD HEIM, HealthPartners

GARY KITCHING, HealthPartners

SUE KNUDSON, HealthPartners

SYED MEHMUD, Wakely Consulting

GREG POPE, RTI International (by teleconference)

EUGENE RICH, MD, Mathematic Policy Research

SHEILA ROMAN, MD, MPH, Centers for Medicare & Medicaid Services

CHRISTOPHER TOMPKINS, PhD, Brandeis University

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

T-A-B-L-E O-F C-O-N-T-E-N-T-S (Con't.)

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

(8:31 a.m.)

MR. AMIN: If it's okay with the Chairs, I just had two quick announcements. So good morning, everyone. Thank you again for all the hard work yesterday. I know that it was sort of a long day going through those two measures. We look forward to sort of finalizing the discussion on the second measure.

Based on the a few emails and conversations I had since last night until this morning, I wanted to clarify two points that sort of process related and sort of reflect a little bit of our discussions. The first is I think many felt a little bit uncomfortable with the fact that there was sort of a split vote on the last measure. Specifically I mean we obviously haven't gone fully through the measure completely at this point but we wanted to clarify that the process of gaining consensus on these measures

doesn't just end today. After the committee has its deliberations on this measure, it will go out for public comment. Our NQF Members broadly will vote on these measures. And the committee will come together during the comment -- after the comment period to review the comments from the membership and the votes to reflect on its conversation, vis-a-vis the conversation of the broader quality community.

So we will think about what other kind of inputs we may need at that point but this is not the committee's last chance, if you will, and it is not the last chance of these measures in any way to move forward.

In addition, the committee deliberations will then go to our CSAC, which is intended to think about these issues on a more macro level. That will address some of the issues that some of the committee members described in terms of the SES potential issues and the dual eligible status issues in terms of risk adjustment and ensuring that this

committee is consistent with the committees across various different projects historically.

So it is not to say -- I say all that to say that the process of gaining consensus takes a bit of time. It doesn't just end today and it will be informed by our colleagues across the NQF membership and the public at large.

I mean obviously, some of those comments already have provided during the committee deliberations and in preparation to the committee meeting. So that is the first issue around how consensus is actually defined and when we actually know we have reached some level of consensus.

And I will say after the CSAC, it will go to the Board where the Board will also discuss these issues. And some of you will remember that in terms of the all cause discussion.

The second is an issue that I

heard yesterday, I got an email about it and also heard it this morning around a little bit of frustration with the fact that it feels like we are giving measure developers input that is much too late in the process. Why does NQF kind of sit where it does? Is there any way that we can provide input earlier in the development process so that what we get at this of whole enterprise is a better measure? It is better for developers. It is better for the steering committees. And it is better for all of us as an enterprise.

So that is a bigger question and that question involves sort of our forward-looking thinking in terms of our own process improvement efforts of NQF.

I will say that we, as Staff, recognize that this is a challenge. It doesn't really help to create a quality check as anyone knows in any manufacturing process it doesn't help to have a quality process. That is at the end of after you have developed

a vehicle to then test to see if it works. You want to have some process that is iterative.

And so we are also trying to maintain our role as a neutral convenor and an endorser, not as a measure developer. And so we are working with CMS, with ONC, and other measure developers in our broader community to think about how we can, as Helen is describing it these days, interdigitate, which we now believe is actually a word, this whole process so that it is iterative and, in some way, we actually have a better process at the end.

It is only to signal that this is effort that we are working on. We don't have a proposal yet, but it is something that we are actively working on. So I don't know that that is particularly satisfying for those that raise that concern but that is, at least, our forward thinking on that issue.

So I just wanted to close the loop on two issues that were raised, the consensus

issue and how it can better provide input to measure developers broadly. And so I will turn it back to the chairs to continue our discussion from yesterday on the measures.

DR. NELSON: Thank you, Taroon.

Well good morning everyone. Thanks for being back. I know that there is a lot of energy and deep feelings and thought about the rest of this discussion on the per capita cost per year measure.

Yesterday was were able to get through one measure and had great discussions. And we now have two more topics for this second measure, being the feasibility and the usability. Dan had mentioned that he and -- is it correct that Dan and Dolores had spoken? Dan and Lisa had spoken. Okay, Dan and Lisa. Okay. And they are going to do a point/counterpoint when we get to the issue of usability.

So I think what we want to do now is pick up feasibility. And for opening

comments, we will have Dolores. A heads up on the rest of the agenda before I go into feasibility and usability.

The original agenda we had three more fairly significant topics: harmonization with other measures, specifically the one developed by health partners which will come after this discussion; and then risk adjustment; and attribution by the end of the day.

So this is the most important is to have a thorough and good and final decision on this measure. And then I think there will be a brief set of remarks from the Acting CEO for about ten minutes or so. And then we will go to the harmonization topic.

So any questions about the setup for today? We will finish as scheduled at 2:30.

Okay, so hearing none, feasibility, Dolores.

MS. YANAGIHARA: So like the other

measure, I think that most people felt that this was highly feasible, using electronic data. It is data not necessarily as part of care but as a byproduct of care in terms of billing and is available electronically. So I don't think that there was any issue there. There were a couple of concerns around the availability of data for other populations, although if this is being approved just for kind of the senior population, that may not play in as well or as much. And as always, kind of concerned about the cleanness of the data, that the quality of the measure depends on the cleanness of the data but that is always the case with data.

So there wasn't any -- I don't think there were really any major concerns that were raised and so that is kind of the summary of feasability.

DR. NELSON: Thank you. So let's open this up for discussion and comment, feasibility.

(Pause.)

DR. NELSON: Did Dolores say it all for the group? She didn't quite say it all. Jack, thank you.

DR. NEEDLEMAN: I just wanted to reinforce a point that Dolores made softly, which is I think it is perfectly feasible for CMS to do this. Any physician, any group should be able to get the data they need. They will be completely dependent upon CMS to do it.

Likewise, we heard yesterday when we were talking about the other measure, that an insurer could implement it. But to the extent that one of the prime targets for this measure and other measures, the other measure yesterday are the providers without access to all the claims data from all your patients. There is no way that a provider can independently implement this measure, which means that it is a measure that is important to get commitment from those who have the data

to make the data available to those who are going to use it.

DR. NELSON: Jack, to clarify, when you said all claims, beyond Medicare or just in Medicare?

DR. NEEDLEMAN: Well the whole risk adjustment, for example, the HCC risk adjustment completely depended upon knowing all the diagnostic codes that were put in anywhere on an ambulatory claim or a prior hospitalization for these patients.

If you are one of Dan's docs, you don't have access to all of that. And if you are one of Joe's hospitals, you don't have access to all of that but it is inherent in the risk adjustment model.

DR. NELSON: Thanks for bringing that point up. David?

DR. GIFFORD: In the spirit of a warm-up, it is more of a comment I was thinking about this last night. Work we have done with physician groups in the past, most

physician groups can't tell you who their panel of patients are. They don't even know who they are. They can't figure it out. So, I'm not sure how you can do attribution when they can't event figure out who they take care of themselves.

So from a feasibility standpoint, that is just sort of a -- I find that sort of almost that we are talking about an efficiency measure when they don't even know efficiency, who they are taking care of. So that is just a broader comment on this issue. So in some sense while I think I was really critical attribution and concerned about it as I thought about it, the docs, this is probably better and helpful to them, even though it is not perfect. And so I have sort of switched my thinking on that. So a long way of saying is I think it is very difficult and feasible to do the attribution but it is better than anything else that is out there.

DR. NELSON: Thanks, David. I

think the first point that you raised is we will hit again in the usability discussion.

Okay, getting ready for the -- winding up for the pitch. Are we ready to vote on this? Is there need for public comment before voting on this topic?

Okay, so we are getting ready to vote. Evan is going to tell us when we have an opportunity to use our clickers.

MR. WILLIAMSON: We will now vote on feasibility. You have 60 seconds, beginning now.

(Pause.)

MR. WILLIAMSON: And we have 19 high; 5 moderate; 1 low; and zero insufficient.

DR. NELSON: Okay, not as close as scientific acceptability.

And so the final topic then, usability and use for this measure, per capita cost measure.

Dan and Lisa indicated that they

were going to work together. Is Dan suiting up?

DR. LATTS: I don't know. He assigned me the negative. So I can start with that. So the usability -- and he will hopefully jump in with the positive when gets back.

So the four things we are looking at for usability is accountability/transparency; improvement -- has there been progress demonstrated or do we think this will lead to improvement; unintended consequences; and then measure deconstruction -- can those who are using the measure understand how it was put together, put the pieces together and then use that to lead to improvement.

And so I think it is actually quite easy to argue that this usability for this measure is low.

So from an accountability/transparency perspective, this

is just mushed together of all of your claims costs. So the accountability to the individual docs, since it is being reported on a group level, is going to be extremely difficult. As far as transparency, from a public transparency standpoint, as far as I can tell, there is no plan to make this transparent to the public. So that is something I definitely actually have concerns about.

As far as improvement, has their progress been demonstrated? This will be a new measure so it hasn't been out there, yet. So we don't have any evidence of this.

The unintended consequences I think could be quite real. And maybe almost in a nefarious way you could think maybe it is intended, depending on if you believe there is a government conspiracy of one way or the other.

The way around this would be to ensure that your more expensive population in

Part A and B were not in Part A and B for all of the year. So all you would have to do would be to shift them into MA or some other arrangement for a month and they would be out of your population.

So I don't think it would take a very sophisticated group to figure that out and go ahead and do it.

I think it would be a little harder to not spend certain costs inappropriately, especially given the quality measures. But I think to shift them into another product line for a small period of time would be not that difficult and actually quite easy for a sophisticated group to do.

As far as the measure deconstruction, for a group to figure out what to do based on this measure alone I think would be quite difficult. I actually, you know to sneak into Daniel's area, was pretty impressed with the reporting. So I think that there is a lot of good information in the

report and I think you could take a lot, take some action from that, although again I think you would need some sophisticated data people and analysts within the group. Otherwise, you are going to get lost in the 40 page report and not know exactly how to act, especially given that it is coming at a group level as opposed to an individual physician level.

So I will stop there and hand over the difficult job of defending it to Daniel.

MR. WOLFSON: Just to set this up a little bit, when you do a counterpoint, it is not necessarily what I believe. So you have to play it like you are role playing. So I just want to make sure that when you attack me, you say I didn't like the counterpoint. You don't attack me.

So I just want to also -- I don't know if you went through the scoring.

DR. LATTS: I did not, no.

MR. WOLFSON: Okay. I'm sorry I left the room. God, you guys were so fast on

feasibility it is ridiculous.

So anyways, the scoring was really interesting. Under improvement, ten people gave it an I for insufficient. I was really surprised by that. And you gave it high marks in accountability. Thirteen people gave it a high mark for accountability and transparency. Unintended consequences was eight gave it insufficient in that category as well and seven gave it medium. And then measured deconstruction, nine said it was -- gave it a high mark.

So I am going to take the contrary view. And I know this will irk people totally. But I actually think that this measure actually can drive change and responsibility model for physicians. The assignment of physicians in a primary care because they are the high utilizer will actually define primary care in a responsibility model that I think is actually very positive. And the specialist will either

have to really own up if they are doing primary care or stop doing primary care, which I think it would be an interesting thing to think about in the marketplace. So I actually think that this notion of how can we make specialists responsible for care is something that I think is actually a positive thing. If specialists want to do primary care, they should.

I worked at Fallon Clinic. Our pulmonary doctors were also primary care doctors. Our cardiologists often acted as primary care doctors and did that quite well. They are internists to begin with we always say. They have general internal medicine underneath them, so they should be able to do primary care. And society actually needs specialists to take that role because we don't have enough primary care.

The 25 provider thing people have talked about but I actually think it coincides with the notion of the law. And I think that

is where we are going. So I don't think it is an inappropriate mark.

I think the attribution has problems but again, I think that there is a positive side to that.

I do think that if you look at Exhibit 9, I think, there is some detail to be able to look at to improve. It might not coincide with how you think about your costs but there is enough detail in there to improve.

And there were some comments about well, what do you want people to do, just provide less? Well you know, first of all professionalism kind of takes over. You know people don't kick people out of panels just to game the system. I think professionalism kind of takes over and guides that along. So I don't buy the gaming. And physicians don't have time to sit and game in this day and age. I find that offensive.

The measure --

(Laughter.)

DR. LATTS: That was a counterpoint.

MR. WOLFSON: And she's a physician. I'm just an MPH.

(Laughter.)

MR. WOLFSON: I'm trying to play this pretty cool.

The measure -- you know there has been talk about the measure is not in broad use. Well, that is not true. I mean, this is a commercial; people in the commercial world do this all the time. And we were talking they were doing it in Michigan. So I think this might be new to fee-for-service and it might be new to Medicare but it is not new to the commercial world. It is certainly not new to people in prepaid systems like HMOs and so on. So I don't buy that.

And I do think it is relevant to policy, where we are going with this law, that we are really kind of matching up to. And I

think it is where the country is going.

So I, in my role as counterpoint, wanted to make those points. The unintended consequences, every measure that looks at any cost or utilization as a tremendous unintended consequence but only can be tempered by quality indicators and professionalism and wanting to do the right thing by patients.

So that is my shtick and I will stand behind it only for this period of time.

DR. NELSON: Thank you, Dan. I think the point/counterpoint has been appreciated. Very good.

Let's open up for general discussion and comment. Bill, I'm sorry. I think you were first, Nancy.

DR. GARRETT: So I have a question for the developers. So we have the exhibits of the reports. Are there plans to give more detailed data so just at the claims detail level to providers as well?

MR. BALLOU: So we continue, as we

produce each cycle of reports to try and improve on the level of detail that is provided. I believe that the Exhibit 9 that is being referenced is probably from last year's reports. And so there are additional categories and breakouts. We have heard from people that they want to hear more about Part B drug break outs. They want to hear about any number of miscellaneous sorts of services, ambulance services, et cetera. So we are continuing to break those out.

We are breaking out -- I believe we may have done this previously, evaluation of management services provided by your group, meaning the report recipient versus other groups. And by even for the other groups, the types of professionals within the group who are providing that on down.

We are also going to provide, starting with this cycle of reports coming out later this summer, beneficiary level information. So identifiers of the individual

beneficiaries that have been attributed meant to be responsible to the earlier comment but often times, physicians don't know who their patients are. They will find out, at least under this attribution role, who these patients are, along with information about any of the chronic conditions they may have had that are of focus in this report for chronic conditions; whether they had any of the potentially avoidable hospitalizations that we seek to capture; whether they were hospitalized at all; when they were hospitalized; for what purpose; where they were discharged to; their discharge status upon discharge. So a lot more information than we have been able to give before, in addition to some information on actually the physicians and eligible professionals that we have associated with the group.

DR. GARRETT: Thanks. So I think, giving that information on the attributed population is really key because that is a

huge issue is trying to understand with retrospect of attribution who is actually in your population. So I think that is really good.

Just a comment from a provider perspective, the report breaks out services provided by your group and other groups. And I think eventually having more detail on that utilization outside your system is really important, as we try and really move forward with care coordination. The providers are really blind. I mean we understand what happens within our system but what happens outside of our system without data from payers we really don't know much about that. And it is really important to be able to understand who those providers are so that more of these partnerships can be formed. So I think that is one comment about usability is that at this point with the measure there is really limited ability to understand that external utilization.

DR. ROMAN: Just to respond to that, we have provided, I think since the beginning of the reports, information on how many physicians your beneficiary is seeing. And I think that starts down that road because fee-for-service has been a very fragmented approach to care. And one of the objectives of this program is to make care less fragmented.

And I think physicians are often very surprised at the number of other physicians who are seeing their patients. And I think your point is relevant to that. That is information that we have.

DR. GARRETT: Right. And absolutely within your ACO programs you are giving that kind of data to providers. And so the question if there isn't kind of that ACO arrangement, is it also possible to start providing that level of transparency.

DR. NELSON: I think Bill is next and then Cheryl and then David.

DR. WEINTRAUB: So I remain very troubled by this measure and I think there are problems with attribution that make the accountability very difficult and the potential unintended consequences are large.

Now Gene and I had an interesting conversation right before we started about this measure and how it could potentially be used. And I think that if you have a large integrated healthcare system, really a large accountable care type organization where everything works together and you have primary care and you have the sub-specialties and then you could attribute to the healthcare system as a whole, based on the number of primary care physicians, you may be able to make some sense of it. But we are not really organized that well. We are not organized that way.

Yes, physicians would be surprised at all the other people that are seeing their patients but if that is all that was going to be used for, that would be sort of

interesting, you know sort of interesting to know something but I think that the potential for this to be used in abusive ways actually is very great.

I don't think we can attribute to primary care what the orthopedic surgeons and the cardiologists all are doing. And you say well they ought to know. They ought to have some sense of it. They really ought to in some way be responsible to the healthcare system as a whole and what those cardiologists are doing so they are not doing things they shouldn't be doing but they are not trained to do so. They don't have the backgrounds to do so.

And so if you were going to have -- if things were like at Dartmouth where everything is every well organized in one healthcare system and you could then attribute to the healthcare system as a whole how they are doing, fine. But the world isn't organized that way. And the potential for attribution to come all the way down to the

individual provider is there. That would be a very bad unintended consequence because we can't put on the backs of primary care physicians and we can't put on the backs certainly of nurse practitioners what the cardiologists are doing.

Now should the cardiologist and the orthopedic surgeons be responsible and make good societal choices and not do things they shouldn't be doing? Absolutely. But that is where at least the way we are organized right now, that is where the responsibility needs to lie.

DR. NELSON: Cheryl.

DR. DAMBERG: Sheila, I wanted to just probe a little bit around this care coordination in sort of a siloed approach. I wasn't sure whether you and your team had considered a multiple attribution approach. Because it seems to me Medicare has this opportunity, you know the struggle in fee-for-service sort of nobody is accountable. Right?

And so I was wondering if you had explored the joint accountability and tried to come at this from a much more patient-centered focus looking across the entire episode, such that you are giving this information back to all providers who touched that particular patient within an episode.

MR. BALLOU: Thank you. It is a good question. We have tested a myriad of attribution rules over the past four or five years. And we have tested several multiple attribution rules, one-touch; what we refer to as multiple even, all costs to all providers who touched; multiple proportional, all costs to providers in the proportion in which they touched them, to proportion measured in a number of different ways.

So I think the answer to that question is yes, we have tested it. There are, as with this rule, there are pros and cons to any attribution rule. So I think how you feel about multiple attribution versus

exclusive attribution really has to do how you come down on some of the issues that were being discussed here.

DR. DAMBERG: Right. But I think if you are largely using this as informational to get people thinking, and talking, and considering how they work with their partners in the community, it seems to me that a multiple attribution approach might be called for. Because I do think that there are these problems with just holding the primary care set of providers accountable in this phase.

MR. WOLFSON: Can you give us an example?

DR. DAMBERG: Of how that work?

MR. WOLFSON: Well just, does it exist somewhere? Is there a model for it? I think it is an interesting idea.

DR. DAMBERG: Well, RAND has done some projects where we have looked at episodes of care and we have looked at multiple accountabilities where if somebody is -- let's

say they are coming in for hip replacement surgery. So they are going to have their pre-hospital care, the hospitalization, and then maybe they maybe the end up in a SNF. And then they are brought back into the community. And so it is partly how you define these episodes.

And my sense is at least within an annualized basis, I think you have some opportunity to stitch together discrete episodes. And so I think that this measure could be enhanced by kind of coming at it from an episode of care. Because really, that is what is going on at the patient level.

DR. NELSON: Thank you. David?

DR. PENSON: So I will keep my comments brief because basically, I completely agree with what Bill said.

I think this has got incredible problems with attribution. And it is really not fair either to the primary care providers who are going to get dinged for the stupid

specialists and how expensive we are. And frankly, I am not even sure it is fair to the specialists because I think, Dan, speaking as a specialist, I don't want credit for taking primary care and I don't want to do it. And I think that basically I understand what the intent of the measure is to get rid of this fragmentation but I think it is wishful thinking that it will do so when in fact I think what it will do is it will cause some really bad unintended consequences.

I could go on, but I won't. Bill basically spoke my mind. So we are of the same thought.

DR. NELSON: So we will hear from Tom, and Brent, and Jennifer, and Larry.

DR. TSANG: Hi. So again, I think the attribution is a problem for me because of several things.

I think the measure itself is going against what we are trying to do with the Affordable Care Act and with the rest of

the system. As Dan brought this up yesterday but MGMA released some data about of the 70 percent of the 600,000 healthcare providers, it is 70 percent of those healthcare providers are in practices of ten or less and that was about ten years ago but it has gone down to significantly to I think about 50 percent. But that portion of the disappearance of these ten or less are being bought up by large integrated delivery systems as coalescing with larger networks.

So the system itself is now coalescing and integrating into larger practices but yet we are focusing on the individual provider level data. And so what I worry is that on the other side of the coin are the quality measures, the clinical quality measures. There is an attempt by CMS and by quality measurers to think about larger group attribution for clinical quality. So with the promulgation of the GPRO Tool with Meaningful Use demonstration allowing for group practices

to actually attest for group clinical quality measures. So we see this other trend in really attributing quality to really larger groups.

So here is this measure that is really focusing on individual provider attribution but then on the flip side, we see all these inducements for larger practices.

So I kind of see is there going to be a problem as you overlay this with what is going on with the other part of the ecosystem.

DR. NELSON: Thanks for those comments, Tom. To clarify our comment, I understand I think the point that you are making, that we are attributing the patients in this program based on a plurality of services, primary care services by individual doctors within a group. But then the measure is applied at the group level. So it is both built on individual encounters but then gets applied at the group level.

DR. ASPLIN: Yes, that was one of

the points I was going to make. If you look at the value based modifier, in 2015 this tracks perfectly with the GPRO Tool because it is going to be groups of a hundred or more. This will be the denominator. So I think that roll up is consistent with how they are approaching the rest of the value-based modifier.

I think Cheryl makes a good point. I would argue, though that for the accountability on the specialist side who aren't acting as primary care, we are going to have to rely on their value-based modifiers whether that is through bundled efficiency and resource use or other registry approaches, depending -- and they are going to unfold by specialty. And I don't think that is a reason not to do the macro measure that we are discussing. And I would respectfully disagree with David and Bill that if we wait until the system is perfectly organized to manage total cost care, we are going to bankrupt the

country before we are organized to manage total cost of care.

And we have 260 programs now that are in either MSSP, Pioneer ACO, or the Transitions program and this is the attribution model that they are using. So there is consistency.

Now, I am not making the argument that just because it has been done before means it is okay to do again. Okay, there is a process here and we are asking good questions about it. I agree with Jack's point about the nurse practitioners, MPAs. I think they should be considered eligible professionals in the attribution models. Those models are certainly going to evolve over time.

I would argue that this is consistent with what we are doing at a group level. If we are not organized in ways to manage total cost of care and deliver on the triple aim for populations, we got to get

there. And there is a healthy tension between the discomfort of being held accountable because you are not organized to be able to deliver and being held accountable in ways that are completely unfair. And I think this strikes the balance for where we are in 2013. And I am going to vote for it.

DR. NELSON: We have quite a long list here. And let me try to get the order roughly right. Jennifer, Larry, Lina, Cheryl, Andrea, Lisa, and Jack. And if anybody wants to drop out, that's okay.

(Laughter.)

DR. NELSON: If anybody wants to add in, that is okay, too. We do want to hear from everybody. Jennifer?

MS. EAMES-HUFF: Yes, I have a couple of questions for clarification. The first one is on the use of the measure. Right now it is intended to be used in the confidential feedback reporting in the value-based payment modifier. Are there any plans

in the future to use this in public reporting?

DR. ROMAN: Yes, I think that there is a Physician Compare site that is coming up. I think it is actually up now with just informational types of data on it. And they will be adding to that for public reporting and eventually would plan to draw from this program. We have no specific dates at this point as to reporting of the value-based payment modifier, per se. But you will see in the upcoming rule what their plans will be for 2014 on quality data that they will be reporting.

MS. EAMES-HUFF: Okay. Thank you.

And then the confidential feedback reports that go back, are they reports that are done at the group level or do the individual physicians get reports on their performance?

DR. ROMAN: And Jeff may want to respond to this as well. Since the initial legislation in 2007 that asked CMS to provide

cost data to physicians, given that physicians are the largest drivers of cost in the system, the reports have been iterative. And we have been, over the years, increasing the number of physicians who have gotten reports. And there has been a combination of individual reports to group reports.

And at this point, for the setting of the value-based payment modifier, we will shifting to group reports so that the attribution model that we are talking about is at the group level, even though it does use individuals in order to assign beneficiaries. It is the group that is accountable for those beneficiaries.

And I would point out that the measure is a whole person care measure. And that the groups are in the best position to impact the care that they receive, the coordination of the care, but probably even more important than the coordination of the care, the access of the individual to care,

and the potential to avoid unnecessary emergency room readmissions and potentially hospital admissions. But certainly, I think an approach to access the groups can effect.

MS. EAMES-HUFF: Okay. So my comment on that is I can appreciate starting with group level reporting. And you have to start somewhere but I think we have seen, particularly in the quality arena, there is a lot of variance at the group level that gets masked when you look at the individual doctors. So I also think the confidential feedback reports that go to the different practices there would be value having it at the individual doctor level. I have seen results where doctors in the same practice practicing across the hall from each other have widely different results.

And so when you move to public reporting and looking at consumers who are using this information to select doctors, they primarily look at their individual doctor, not

at the group, even though they do get group care. So, I am not saying one or other but I think there is a place for both of them.

DR. NELSON: Larry.

MR. BECKER: So to me, it doesn't matter what business any of us are in but we can't improve that business unless we actually measure it. Because otherwise what happens is we work as we can. We do everything we can possibly think of. We have no measures to know if we are doing any better. And so Deming would call that tinkering. We are just tinkering with the system.

And so it seems to me that we need to start with something. We need to put it out there. It is certainly not perfect but I think the approach of putting it out there to physicians to let them see that, I think Paul Tang would say there is a quality dividend by doing that because people will look at their own performance and they will naturally want to do better. So I think there is value to

putting the measure out there.

And then as I listen the comments around the room, one of the themes that I think is a latent them that we are hearing is healthcare is a team sport. And so I think what we are hearing is beyond this measure of attribution to primary care is, and I think someone said that earlier, and that is we need the measure that enables the team, whoever that team, that virtual team to be able to look at their performance and their contribution with the patient as the North Star. And so when I look at the patient what happened? How can we work together? Because I think that is where the whole ACO concept is going. As a patient, that is what I would like. I want my primary care and my oncologist and my cardiologist, I want them to work together and I want them to know what is going on with me and not to cross paths and one do one thing and one do another and I end up with a medication problem.

So I think it is incumbent upon us to start here but to, as quickly as we can, move to group kinds of measures so that everybody can improve their performance with real measures and real guidance.

DR. WALKER: Thank you. I will keep my comments short because Larry and Brent said many of the remarks that I wanted to add to this conversation.

You know yesterday we talked a little bit about how perhaps this measure is a little backwards looking and not forward looking but in listening to this conversation, there is an element of this measure that is actually forward looking, in my opinion. We are moving towards integrated systems where we care about team-based approaches and this measure is attributed to the group. This seems to be moving us in the right direction. And I think it is okay to ask primary care physicians or group practices to be accountable for the care of the patients. And

I just want to say that we shouldn't let the perfect be the enemy of the good.

DR. NELSON: Thank you, Lina. Cheryl?

DR. DAMBERG: Yes, I just wanted to go on record that I am not opposing this measure. I think it is actually a very good start and I would encourage CMS to continue their exploration to try to move toward joint accountabilities and really helping all the players because my sense is as well we probably generally don't like fee-for-service it is with us for probably my lifetime. So we have got to figure out a way to make sure that all providers really understand that they are in this together.

DR. NELSON: Andrea?

DR. GELZER: Yes, I agree with Larry and Brent that on the concept that we need a measure like this and we need a measure like this as soon as we can get a measure like this. But having said that, when I look at

this measure, I think it is fundamentally flawed from a validity perspective.

I mean we have experience with shared savings arrangements where we provide, on a monthly basis, a sortable database to all our large groups, which are typically hospital systems with physician-owned practices. And they use that data and they find that data useful to manage these contracts for two reasons. One, they can control leakage. They can see what their leakage is. And secondly, they can then work with the physicians and other providers that are in their groups to manage cost spikes. And so that works and we need to move there and we need to move there in Medicare.

But the attribution model in this measure doesn't really bother me because those groups, as long as they have the data can control, control at that group level. But it is really the potential, and I do think there is potential, of gaming. I'm happy that the

duals are included but I see that there is real issue with the way they have been included and, as I said yesterday, the way with the calendar year measure combined with a Medicare Advantage Exclusion, there is real potential to game this measure and not do what it was intended to do, which is monitor of the cost of the care that these physician groups are providing.

DR. NELSON: Thank you. So Lisa, Jack, Dan, and Matthew.

DR. LATTS: So I actually am also in favor of this measure on a global perspective. I think there are a lot of problems. And when we at WellPoint in my pervious life did research looking at group cost and group quality, the intragroup variation was actually greater than the between group variation. So I think it is very difficult to say this is the cost of the group overall.

And I don't disagree, Bill I think

was the first one that made this comment, that it is very hard to hold primary care physicians accountable for the spend. But if we don't let them know the spend, how are we ever going to get control and get a handle on our spend overall? So you can't be responsible for what your specialists spend but wouldn't you like to know which specialists are spending a fortune and which aren't? And then you can either modify your referral policies accordingly or at least try to understand the underlying reasons behind what is going on.

And so I think that to provide this information that there can never be harm in information. It is what you do with that information that is the problem. And so do I think this is a perfect measure? No. Do I think there is lots of room for improvement? Absolutely. I would love to see Part D included. I think to not to have pharmacy costs in here is crazy. I would certainly

like to have partial year included and be able to adjust accordingly. But I think we have got to start somewhere.

DR. NELSON: Thank you, Lisa. Jack?

DR. NEEDLEMAN: I have been critical about a number of elements of this measure. I actually think one of the things the measure gets right is its focus at the group level and starting at the large group level, where much of the care is in fact going to be, and referrals are going to be within the group. Are some of the attributions going to be wrong? Yes, the folks who get hit by buses or who have heart attacks while they are vacationing in Florida are going to have a lot of costs attributed to a group in Minnesota or New York that they had absolutely no control over. But those are going to be a small portion of these cases.

I agree with Jennifer and I agree with Lisa that there is a lot of within group

variation that the groups need to deal with. And I think a measure that is at the group level that provides group data at this point will allow us to figure out how to deal -- allow the groups to figure out how to deal with that well, the within primary care doctor, for instance, and frankly, the referrals. And the measure provides time for the smaller groups that are making referrals outside of the immediate group to figure out how to manage their referral networks.

And I know in California we have a number of large integrated primary care groups that are being very selective about which specialists they are allowing the group to basically refer to. And that is in part because they have data like this about which folks they think are delivering higher quality care and delivering it more efficiently as the integrated primary care group accepts capitation.

So, I think the individual level

is not right at this point. The group level feels about right starting at the large group level, where the integration is probably better, makes a lot of sense and gives time for the smaller groups to figure out how to do it. As I said, critical about a lot of elements in this measure. This is not one of them.

I would like to see a commitment by CMS to deal with improved risk adjustment. The purpose of this measure is to allow for groups to manage discretionary care appropriately. Where does the spending have value? Where doesn't the spending have value? But a lot of the care that is included in terms of these large costs, things like SNF care and some other forms of care, there may not be a lot of discretion about whether the patient gets that. And the risk adjuster is supposed to fix that. But I would like to see a lot more work on how these risk adjusters do in predicting specific subsets of costs and

some refinement of the risk-adjustment methodology to better predict individual components of these costs so we know that the discretionary care is what is winding up in the residual.

Do I think we have to wait until that happens to say the measure is okay? No, I don't think we have to wait.

DR. NELSON: Thank you. Dan?

MR. WOLFSON: This is a great conversation. I think we are talking about what we want the healthcare delivery system to be and I think that is great.

I have a few comments. On the individual level, this is coming from a person that works with the American Board of Internal Medicine. I think we are at the right level at the group. And I wanted to know from the researchers whether on the individual level there was stability of those measures. I mean we are not talking about all payors, we are just talking Medicare and I am worried about

those cells being too small. So I would like to hear how those individual measures held up from a psychometric point of view.

A lot of issues about gaming and I wanted to know from the researchers or NQF was there an audit planned for this so any gaming would be caught. And just the threat of an audit usually calms down gaming. Nobody wants to be thrown out of the bin.

And I also would like to see pharmacy put in. I know at only 60 percent but let's see what happens to people who are on Part D and who are not and see what their total spend looks like. I think that would be an interesting question to pose.

So anything on audit and individual stability of the measures when you get there? I am not advocating it, by the way. I think the group is the place to go. I would advocate not doing it but I wanted to know for just the information for a medical group, is that good data or not.

MR. BALLOU: So I can address the stability, which I am inclined to address in terms of reliability testing that we have done at the individual level. And then I will ask CMS to address the gaming and audit issue.

As you may recall, we have imposed a minimum T size. Even for groups here you need to have at least 20 attributed beneficiaries. When we imposed that for individuals as well, we get reliabilities that we initially found to be surprisingly high. It is a Medicare population but you still do have a significant majority clearing the 0.7 threshold that we have been using for reliability.

So you do have the reliability at the individual level. However, let me back up. As I told Cheryl before, we have tested many rules. And I do not believe we have yet tested reliability for this particular attribution rule. We have tested it for a more limited set of E&M codes, exclusive

attribution in one step.

So I think the takeaway there is that we have evidence for exclusive attribution broadly defined. These can be more reliable than we originally thought at the individual level. However, it hasn't been tested on this current rule and we have been responsive to physician feedback, which has been echoed in many of the comments around this room that certainly to begin, physicians do expect and prefer to be assessed as groups.

DR. ROMAN: I think on the audit side of things, I think the Agency is clearly interested in preventing gaming. And we will be looking at this data that there are plans in our innovation center to begin looking at some of their value-based payment programs. I am not fully aware of a specific audit plan in the innovation center for this specific program but clearly we will be having our contractors look at a variety of impacts that occur with the institution of the system,

particularly as we move through the implementation years. And clearly, that is part of the approach of the agency in moving slowly and carefully in bringing the value-based payment modifier of which this measure is foundational on the cost measure side. We haven't talked much about the quality measure side of the value-based payment modifier here. I think that is reflective of the fact that the Agency is concerned and will be monitoring.

MR. AMIN: Just to quickly add from the NQF perspective on this issue of the audit, NQF, as part of measure maintenance, any stakeholder that recognizes any -- has evidence of any unintended consequences to patients can submit that evidence to NQF and it will trigger an ad hoc review for a measure. So that would, obviously, be in play for this type of measure as well.

DR. NELSON: Matthew, then Nancy.

DR. McHUGH: So I think that I

would support this measure if it was what it says that it is. But I have concerns, I think as we have discussed, that the attribution issues and validity issues really kind of killed that. And it seems to account for some providers in some context and account for some beneficiaries and services. And it is too conditional to really make it a valid measure of what I think the provider community and then ultimately patients would want to see in terms of information around this measure.

I also have concerns about not per se that socioeconomic status or dual eligible status is controlled for. I just don't think that the developers have made the case here in this instance. So I think more evidence of the rationale for that is necessary.

DR. NELSON: By too conditional you mean? Matthew?

DR. McHUGH: The exclusion criteria. So for instance on the provider side, the attribution rules excluding a subset

of providers from the first step. So NPs and PAs, for instance. And then the kinds of issues that we talked about in terms of the beneficiaries that are included and the kinds of cases and the information on those cases that would be included, deaths and those kinds of cases. I think that is valuable information.

DR. NELSON: Thank you. And Nancy?

DR. GARRETT: I just wanted to respond to Dan's question about an example of using multiple attribution rules.

So in Minnesota we have a project, my colleagues from Minnesota can chime in, called Provider Peer Grouping, which is a state-level project to do provider profiling and measurement. And there is a raging controversy about whether to do multiple proportional attribution or single clinic attribution. And kind of the current approach is that multiple attribution. And it is

really tough because intuitively it really creates a usability problem. It is a lot harder to understand. And so there is a lot of debate. So I don't know that we have a successful example for you but there is a lot of conversation going on about it.

DR. NELSON: Thank you, Nancy.

I don't see any more cards up. I have seen some cards go up and down. This would be a great time for any final thoughts on this final topic of usability and use.

(Pause.)

DR. NELSON: So we will have a vote on this and then we will have a public comment and then we will have a final vote for this group at this time of yes or no.

So usability and use, Evan is going to open up the polling platform.

MR. WILLIAMSON: We will now vote on usability and use. You will have 60 seconds, beginning now.

(Pause.)

MR. WILLIAMSON: We are waiting on one more. If everyone could please point again at the -- there we go. Four high; fourteen moderate; seven low, and zero insufficient.

DR. NELSON: Thank you. So we have covered the criteria one by one. And now it is time to have an overall discussion about this measure and to hear public comments. So why don't we start with public comments and then we will have a final opportunity for this committee to weigh in.

MS. TIGHE: Operator, is there anyone on the line who would like to make a comment?

OPERATOR: To make a comment at this time, please press *1. Once again, to make a comment at this time, please press *1.

There are no comment or questions.

DR. NELSON: Okay, thank you. So we have an opportunity for more discussion deliberation before our final vote. The final

vote is simply yes or no or one or two.

Andrea?

DR. GELZER: Yes, I just wanted to say that I do believe that there are real flaws in some of the stuff in this measure that we have discussed. But I do think it is critically important that we have these cost measures in use and people start to both deal with them and gain experience dealing with them and understanding them and gaining additional data from them.

And as I think was said by the chairs yesterday, you have to decide is this so fundamentally flawed that you can't vote for it? I would just ask CMS, I would implore CMS to look at this. Whatever gaming that goes on, people are going to have to understand -- the first year this measure goes into effect I don't think there is going to be a lot of gaming. I think it is when people start to see their results coming back. So I would just hope and expect and really demand

that CMS makes this measure better and better and does better, you know make sure that the risk adjustment and the socioeconomic factors considerations are improved as the years go by.

DR. NELSON: Bill and then David. And try to, if possible, make new comments, rather than reemphasis, if possible.

DR. WEINTRAUB: And I will do just that. I will not say what I have said before. I do very much believe in economic measures. I very much believed in the measure that we voted in favor of yesterday. My own research is concerned with that measure and I am convinced that it is going to measure real things.

I think there is a danger in approving a measure, however, that I think is so fundamentally flawed that it is not going to drive the whole process forward but could set us back. I remain unconvinced that we can handle the problems of attribution with this

measure in a way that will be meaningful.

DR. NELSON: Thanks, Bill. I was just giving a guideline. It is not a protocol.

David?

DR. GIFFORD: Question to Helen. In the past when we didn't like the yes or no vote, we added a third time limited or anything. That is no longer an option?

DR. BURSTIN: Well that is only for measures that have not been tested. This measure has been tested so it needs to be yes or no.

DR. GIFFORD: I can't add anything original to the discussion that has been had so far.

DR. NELSON: Okay. Perhaps not everything has been said but many things have been said.

So time for vote. One, yes. jack?

DR. NEEDLEMAN: I oppose the last

comment about provisionality. There are two things that are happening here. One is CMS was obligated to develop the measure and implement it and it is doing it. So now the question is what role does NQF endorsement play in the process? And the issue, in part, is is it good enough to merit endorsement and the endorsement with all the comments we make encourage the kinds of changes that are needed to make it better.

Well apparently, it is the withholding of an endorsement from a measure that is clearly going to be used, a clear signal to people that the problems that have been identified in the discussion with the measure need to be addressed in the future and come back to us as you keep working on it.

So I think to me that becomes the issue. And think about this. It is not about whether CMS is going to use it. They are going to use it. They are obligated to use it. The issue is whether NQF as an

organization and us, as a steering counsel, are signaling how much room and need for improvement there is in this measure. And on that basis, I am going to vote no because I think the attribution that is particularly important to me and some of the other issues that have been raised to others are sufficiently important that I want those fixed before it becomes an NQF-endorsed measure.

DR. NELSON: Helen?

DR. BURSTIN: Just one brief comment on that. I just want to really encourage you again, I know it gets very complex when you know what the intended use is. You know it is already on the street and you know what is going to get used for it.

Your role really is the science here. You really need to vote on overall suitability for endorsement based on our criteria for endorsement. All this externality stuff we can comment on and put in the report. We really bring you together as

experts and multi-stakeholders to bring the science to the table. So, please vote on the science.

DR. PENSON: Can I make a comment? Sorry, Gene.

So Helen, I hear what you are saying and I would like to sort of live in that vacuum but I am having a hard time with it because we all know how this particular measure is going to play out.

And I share everyone's desire to push the field ahead and recognize that this will push the field ahead but I also know full-well this -- I feel this has got some real problems both from the validity standpoint and from the usability standpoint.

In the end, it is not ready for prime time but it would be nice to push it out there just from the science, say it is not perfect but it is good enough. But this one, in particular, because of its clear applications and it says in the applications

can be used for the value-based modifier, I am having a really hard time saying, okay, the science is not perfect but it is okay because the stakes are so high and I know what the outcome is going to be. And none of us want to say it but everyone in the room knows it. So I don't know how to respond to that, except say I am having a really hard time with that.

DR. BURSTIN: It is really complex. We have had very similar issues with some of these measures that you know where they are going, you know how they are going to be used. But at the end of the day there are other groups that are going to take more consideration about, for example, the MAP that I will really think through, for example how the measures are applicable or not applicable and given programs. We know this measure is going to get used. So I guess the question, one question might be is it better to sort of have it inside the portfolio and potentially work with CMS to have it get modified over

time to be more reflective of a lot of the concerns you have raised or is it a better signal to indicate it is not ready for prime time? That is where you have to weigh -- I mean those four criteria are there intentionally. You need to weigh in your head -- I'm not prescribing to you that 30 percent of your assessment overall endorsement should be validity and 40 percent should be usability. This is your chance to say I have now voted on the four criteria. How do I collectively weigh those four criteria to make my final decision of endorsement?

And again, keep in mind, as much as this feels like sort of a final step, it is quite early in our consensus process. So this measure will still go out for comment. You will have another chance to reflect on it. I don't know what CMS's and Mathematica's capacity is to potentially respond to any of the specific concerns raised today and whether any of these issues are mutable in the short

term to actually make it a measure that might be more -- you know to have less concerns. But that remains an issue.

Yes, Lisa?

DR. LATTS: So that actually was a question I was going to have. Can Mathematica take the concerns that we have raised today and revise it before it goes to the MAP, or address it, or whatever the next appropriate step is?

DR. BURSTIN: Yes, I mean this is a pretty complex measure. It is hard to imagine there is a whole lot you can do without having to go back, retest it, re-put it forward and all that stuff. I mean what the MAP, I believe, put forward was that they would support the direction, depending on NQF endorsement. So it is kind of back to you.

So I don't know. It is going to be a somewhat circular argument, I think. So at this point I think you really need to consider it. I don't know whether Sheila has

any comments or anything from CMS's perspective about how much and how timely some of the responsiveness to some of the concerns raised could be in the course of this project.

DR. ROMAN: I mean I don't think that I am at liberty to get out ahead of the Agency on this but I would say that we obviously take your input quite seriously. You know we purged three or four major issues that have been problematic for the panel and that I would go as far as to say that we will be dealing with these issues, that they would be probably going through our formal rulemaking process. So we are not talking about the next six weeks but that we have clearly heard the problems and understand where the committee feels that there are major flaws and that we understand that we have to look at that very closely. And the agency will need to make some decisions on where it wants to take these recommendations. And I am sure that they will do it.

DR. GIFFORD: Helen is really going to kill me. You asked me to come to this meeting.

In the past, again I know the continuous quality improvement aspect is well adopted by NQF and I do think that each of these panels has gotten better that I have been on over the years.

We have held a vote and had a group conference call, giving some developers some time to address some questions and come back to us. That is one question, is that an option here.

And the other one is, if we vote to approve it, and they make all these substantive changes that they are talking about doing, is that enough that it is no longer -- it is different so it has to come back to us anyways because you said major changes have to come back to us or that only applies once it has gone through the whole consensus? And how do those additional

changes get reviewed in this process? Because these aren't just like minor tweaks. They are substantive changes.

DR. BURSTIN: And many in the room will know this well since we went through a fair amount of this with some of our readmission measures in the past.

So there was certainly an opportunity for the committee to vote today, to put it out for comment and just so you -- you know, NQF had changed the policy a few years ago, so all measures go out for comment, regardless of whether you approve them or not. In the past, we only put out for comment those you approved. Regardless, those measures are going out for comment, you will get, I assume a significant number of comments on these two measures. You will have a chance as part of the follow-up conference call to review the comments to consider whether any of the comments and, frankly, CMS and Mathematica will have an opportunity to also respond to

any of the specific comments directed at them about the measure, in addition to responding to the whole host of issues you guys raised today and yesterday. So at that point, they will have a chance to respond.

We do frequently if a committee feels that based on the comments they have seen, the responses from the developer, you will have an opportunity to revote on that post-comment call, if you feel like that the world has sufficiently based on input from the commenters, input from the developers, additional analyses from the developers to potentially respond back, to some of your concerns. So that is, certainly, very much part of our process.

Does that help?

DR. NELSON: Carolyn.

MS. PARE: I think, Helen answered some of the questions I had. I am really struggling because, as a purchaser, we are hungry for this kind of evaluation

measurement. But having been from the collaborative state of Minnesota and just seeing some of the struggles we have had in coming to an agreement on what these kind of measures should look like, I just know that more broadly, across the country, this is going to have significant challenges and that I don't want to shoot ourselves in the foot by moving something forward that just really is going to get so much pushback that it is no going to go to a place that we want it to go.

I was interested, really, in understanding the process. Because from this point, as I understand it, if we move it forward, it goes to the MAP.

Could you just take me through the process so that I understand what kind of damage my vote does at this point?

(Laughter.)

DR. BURSTIN: That is a very loaded way to phrase that question, Carolyn.

At this point, it is truly, this

vote is about endorsement of the measure. Please take everything else off the table. It is about endorsement of the measure. The MAP did indicate it has already reviewed the measure, I believe, for the IQF. So they have already made the recommendation that they would support the direction of this measure pending review by you. So I assume they will have an opportunity again to reflect back on whatever comes out of this process. But again, it is so early in our consensus process, I mean you will get, I suspect, I am going to put out there, hundreds of comments on these two measures that you will have a chance to sort through, that Mathematica and CMS will have a chance to respond to. So I think there is a lot more to this process than your vote today. Your vote today is important because I think it signals to the wider world, where the concerns are, where you think it should potentially move one way or another but either way, it is going out for comment.

And either way, there is an expectation the developers will respond to those comments and to any concerns you have raised today.

Taroon, do you want to add anything, based on your prior -- where is he? Okay, never mind.

MS. EAMES-HUFF: Can I ask just a question for clarification? So if this measure is not passed by the committee, it still goes out for comment? Is that true?

DR. BURSTIN: All measures go out for comment. And we put in the full commentary of what happened at the meeting, regardless. And I suspect this will be a somewhat close vote. So either way, you are going to have a fairly rich discussion in the report and comments.

DR. LATTS: I just wondered what happened then. So if we vote now, it goes up for comment. Then what?

DR. BURSTIN: All those comments come back to you on that post-comment call.

You review all those comments. You would then have an opportunity to say, based on those comments, somebody could move to say we believe, based on how much comments we got, the response back from the developers, we would like to re-vote. So you have an opportunity on that post-comment call to revote on the measure again, with the additional information if you think, again, there is anything you are going to hear that is going to potentially change your mind.

And again, given the number of comments and the richness of the comments, I think there will be a lot of substrate for you to get through.

MR. BECKER: So this sounds like it is a real struggle for everybody. And so-- and you might kill me on this one -- but what if we didn't vote?

(Laughter.)

MR. BECKER: What if we held and got all these comments and got all the work

and then reconvened and said okay, now we have more information and we have more work done. And then --

DR. BURSTIN: Since you are on my Board of Directors, Larry, I will take that -- we cannot change the process in that way. However, it is very reasonable for your vote for all of you to consider your vote your first vote. And if that helps you all to say this is your first opportunity, based on everything you have heard today, to vote, knowing you will, in all likelihood, have an opportunity to reconsider that vote, post-comment. If that helps, Larry, then think of it that way but we cannot, not vote.

DR. NELSON: So we had a pre-meeting, non-binding vote. We have a vote now. And then we have a vote after. That's not bad.

MR. BECKER: It's a two out of three.

DR. LATTS: Are you sure we are

not in Chicago?

DR. MARCINIAK: So being from Chicago, I would kind of agree with that comment.

You know, the reason why I am struggling with this is not -- because we have had the robust scientific conversation. We have talked about multiple attributions, single attribution and things of that nature. We have talked about the risk adjustment. As a researcher, I feel comfortable with that.

The struggle I am having actually is where we left the validity vote yesterday because we left it as a hung jury. It was split 13-12. And so you look day over day, our votes look a bit different today with 24 hours of reflection. And so when we talk about usability, if it is a hung jury on validity, how can it be useable. And that is where I am having difficulty with this because we didn't really wrestle to the ground what the validity issue is. And we really didn't

change hearts and minds during the course of the discussion of where we wanted it to go. If the measure is not valid, we are not getting to this question. Right? And so we are getting here split. We very well still may be split.

DR. BURSTIN: I suspect you will be. And I would just encourage you to vote.

DR. NELSON: As a further point of clarification before we vote, I think we are going to vote soon, Helen, after we are done voting we have a discussion about harmonization. And I know I am on the MAP Clinician Work Group and I know there is a lot of interest in having public-private payer alignment. And this harmonization discussion around two per capita measures opens the opportunity for thinking about alignment.

And so what I am wondering about is the interaction, if you will, between a measure that we are going to look at that is behind us, the one that is in front of us, and

actually coming up with something better than either that applies to all ages.

DR. BURSTIN: Either way, I think we will have that discussion. And I think that is all very future tense. I think anybody making significant changes to align is something that is not going to happen tomorrow either. So I think again we do have that -- yes. I still think we will talk about it.

DR. NELSON: Okay, let's vote. One, yes; two, no.

DR. GIFFORD: Actually since we are in D.C. I would like us to start reading from Harry Potter and filibuster the vote for a while.

(Laughter.)

MR. WILLIAMSON: We will now vote on the overall suitability for endorsement. You will have 60 seconds, beginning now.

(Pause.)

MR. WILLIAMSON: And we are still waiting on one response. There we go.

Eleven yes, fourteen no.

DR. NELSON: So we have had a great discussion. And why don't we take a break. It's ten o'clock.

(Laughter.)

DR. NELSON: And we can reflect and start on our next topic. Thanks everyone for a really thoughtful and deep discussion. We all know that this is a high stakes measure. It has been given a lot of consideration and we will have a chance to consider it further. Thank you.

MS. WILBON: Let's plan to return about seven after if we can, at least be ready to go by ten after. Thanks.

(Whereupon, the foregoing proceeding went off the record at 9:56 a.m. and went back on the record at 10:11 a.m.)

MS. TIGHE: Okay, thanks everyone. We have Gerry Shea, our interim CEO here. And he is going to lead a conversation with you

all about the future of Steering Committee meetings, I guess.

MR. SHEA: Thank you, Lindsey. Good morning. I did want to impose on you to take a few minutes to talk about an issue that has arisen in our discussions with the federal agencies about these meetings. But first, let me just say a big thank you for wrestling with difficult issues like the one you have got before you and had before you yesterday. Needless to say, these are not easy but they are enormously important. So the fact that you are willing to come and put your time into this is really very, very significant.

I don't know whether you have heard the number but we counted last year and there were 55,000 hours of expert volunteer time in meetings or on webinars not in preparation, not flying to get to the meetings, actual meeting participation, which amounts to a contribution that you and your colleagues make or your organizations make of

about five million dollars to this whole process. So thank you, very much.

I won't take very long this morning but I would appreciate your comments on something. And the situation is that due to sequestration and the budget cuts that flow from that, and also in the wake of the scandalous GSA meeting, I guess it was, in Las Vegas last year, the administration generally, including HHS, has been very tough on in-person meetings under federal contracts. And they have moved as far as saying to us not only do they have a very laborious process of approving an in-person meeting, and I won't bother you with my day job problems, but they have said to us they really want us to move to virtual meetings.

And we are concerned about what the impact might be on the quality of the process or the ability to really even do this or what we are imposing on people. And the bottom line here is we are not going to do

anything to compromise this process. But on the other hand, we do want to be responsive to explore the possibilities of maybe improving our -- or significantly improving our technology to be able to do more virtual meetings.

But just while you are here, and in light of the kind of discussion that you had yesterday and this morning, I just wanted to get your comments on to what extent do you think these kind of meetings can be done remotely and to what extent do you think or under what conditions do you think they need to be done in person.

I should say that the main responsiveness to the concern is as part of our reengineering of the measure review and endorsement process, we are hoping to go to standing committees. And those standing committees will have three-year terms. So once those standing committees get to know each other and get sort of working, then it is

probably more possible to do things remotely. So just as an example.

And we are also consulting with NIH and the IOM people. And well what do you do with these sort of situations? So people have comparable kind of things. And we are going to look at the best the brightest in the global business world on how they do these things, not that they are always so good or so bright.

So I would be very interested in your comments on this and just want to take a few minutes and not exactly go around the table, but I would encourage as many of you to weigh in on this as possible. And this is just useful input to us.

DR. WEINTRAUB: Well first, thank you for having us. And this is a fascinating two days. I always feel that I get much more out of these than I ever contribute to it.

That being said, I think that the in-person is what carries the day on that and

is extremely important. I don't think even with video technology you can get the kind of interchange between people that you get when you are sitting around the table face-to-face really discussing things.

So I think there remains a real place for real people sitting around the table and working together.

DR. WALKER: I think you alluded to the point that it depends really on how well you know individuals on the committee. So I am new to this process and I don't know anybody here but I had a side conversation with Bill on the way to the bathroom after the vote. And so would be lost in a telephone meeting, telephonic meeting. That said, we do a lot of those types of virtual meetings at AARP and are quite effective holding those types of meetings because we know everybody. And what is really integral to this process of the side conversations that go on and there are a lot of side conversations in this two-

day meeting. And if you have a virtual meeting, you can still enable those types of conversations. You just need a little bit more lead time and those conversations would happen over email or somebody would pick up the telephone call. But again, the key components there are that you need to know those individuals on you committee fairly well and you need to have more time. You need to build in a little bit more time before you get to the point where you have your evaluation and vote.

DR. PENSON: I guess we can just go around probably. So I mean there is no one who likes virtual meetings better than me, except maybe my wife and my kids. And so I am all for that but I will tell you that the first comment that was made about a committee knowing one another is critical. Having sat on a number of and continuing sitting on NIH study sections, you know what people are going to say. You know where they sit and that is

helpful.

So I think on the one hand you can say well if we put in a standing committee, we won't need to do this anymore but I am going to give you pushback because there is another element to it as well. And that is, these committees are all about building consensus.

It is one thing when something is fairly straight forward and easy to get to the bottom of. Something like these two issues we have been dealing with the last two days, these two measures are really hard to build consensus around, as you saw. And I think that you would have lost a lot of the richness out of the discussion and you wouldn't have gleaned the information that I think the measure developers are going to need to come back or other committees are going to need to go forward.

So I think you have to consider both how well the committee knows each other but also the sort of controversy, I guess is

the best term or the -- yes, that is probably the best term -- of what you are assessing. And I don't think it will work for things like this is my thoughts.

DR. MARCINIAK: So David hit on some of my topics so I will move in a bit of a different direction. And I work for a large corporation where we believe ourselves to be a very technology-enabled company in terms of TeleSuites, and video, and video on demand. My experience has generally been that most of those technologies don't work very well. We just haven't caught up to a sufficient degree to enable a room of 20 or 25 people to sort of effectively communicate in a way that would allow any of the dialogue that we have had over the last day or two to be facilitated in a reasonable way.

So sort of bridging off with David who suggested and what Lina had said as well, the advantage is what conversations happen in the hallway, oftentimes, not what

conversations that happen in the room to sort of facilitate a dialogue and to help it move forward if you can find the common ground. So I just don't think we are there yet.

MS. EAMES-HUFF: As someone who travels from California, I love the idea of having virtual meetings but I think I agree with the general consensus around the nature of the in-person. But I don't think it is an either/or. I think it is there are different elements and it is a case-by-case basis that would need to be decided if the situation has the right factors for doing an in-person versus the virtual.

The other thing I would say is if you found that you tried a virtual meeting and it didn't go well, I know this extends the time, but there is always the possibility of then calling an in-person meeting if you are not getting what you need from the virtual piece.

MR. SHEA: And Jennifer, if I

could just ask, would you want to identify some of the elements that you think distinguish meetings that could be done virtually as opposed to ones in-person?

MS. EAMES-HUFF: I think the idea of standing committees, where people know each other more is a key component to having. I think also if there are things done in the beginning, before the meeting, if there is like group warm-up that can be done virtually before just starting people in with a meeting is another piece.

I will confess for the bathroom conversations, virtually I am less likely to do, given my schedule. I just know I am stuck here. So it is not just that I am stuck, I like talking to you guys. But when I am in my own office it is just much harder to do. And I think that there is that factor, as well.

DR. DAMBERG: So I have participated in some of the NIH review panels recently and they do that kind of

teleconference where you sort of see a panel on the East Coast and there is one on the West Coast. And so you still get to connect with others sort of in your local area. So it is not completely virtual. So that might be an option.

But one of the things that I have been pondering because I think there was a lot of time spent asking the measure developers a lot of questions to inform people's considerations of how to do the ratings. And it felt to me like there were some efficiencies to be had here that you could do a virtual phone call or two in advance of people coming together such that this meeting might have only had to take place over a one-day period instead of two days.

DR. NAESSENS: Yes, I was going to say I am on a committee at Mayo Clinic where we have a standing committee with representation from Florida and California -- well, Florida and Arizona. And we started off

with virtual meetings about once a quarter. We ended up having too much to talk about, went to virtual meetings once a month. And then realized that we really missed or needed to create the bonding across the group and now have gone first to a quarterly meeting in-person with the monthly virtual meetings and then have actually reduced that down for some cost savings so that we only meet in-person once every six months.

DR. WONG: Well, I'm from the federal government. I feel everyone's pain, especially given the circumstances several years ago with our sister organization and how they conducted themselves. So we struggle with this all the time because we also facilitate many meetings and we have gone somewhat virtual with webinars and things of that nature. It is not a substitution. It is you have to completely rethink of what your objectives and goals are.

In my view, a committee like this

is really a scientific review committee. It is parallel to what you mentioned in terms of NIH, in terms of the grant review process. Similarly, at AHRQ, that grant review process. I think that it would be helpful to kind of contact the folks at NIH and even AHRQ of how they want to kind of handle this sort of aspect of it.

I think that it is important to continue to communicate to others who make this decision about travel rules, about the compromises that you have to take. So if the mandate, whether it is from AHRQ, NIH, or even contractors that support these other federal agencies, I think the message really needs to be clear that we are not quite getting the quality of the review and things of that nature that we are making compromises of that.

DR. DAMBERG: As a confirmed curmudgeon and I am working on my geezer skills and I am almost there, I hate to travel. If I can assign somebody else to

travel for me, I am going to do it. And so I really did not enjoy traveling here but I really feel the face-to-face is just absolutely necessary. And it is more than the verbal communication. There is a fair amount of nonverbal communication that occurs even at the table here. There is no substitute for that.

And if it came down -- I am not volunteering for anybody else but if I had to pay the airfare to come myself, I would do it because I think it is important enough to have the face-to-face contact.

MR. SHEA: Thanks. And I will note that the little bit of literature I have read on this, which is mostly from the business world does always mention as sort of like a big component the sort of non-verbal communication, the body language and so forth.

MR. WOLFSON: The one thing having gone through both these in-person things now for a couple of sessions and a lot of phone

calls, the one observation I have made about the phone calls is I think it suppresses minority opinions. And I think the reason that happens is that you can express an opinion and you don't get any immediate feedback. You think am I just an idiot? Am I the only person on the group that has this crazy idea? And you don't know and it suppresses it. It reduces the likelihood and if you have a minority opinion, it is hard to keep pushing it. Where at this meeting, you get immediate feedback, either verbal or nonverbal about whether you are heading in a direction that people think is a sense. And so I think that is the defect, the main defect I see in the phone meetings, that you don't get minority opinions expressed. And I think one thing we could all probably agree on, sometimes minority opinions turn into majority opinions pretty fast during these discussions. And so I think it is a really valuable function to bring that out in the meetings.

DR. GELZER: I agree with everything said, even with standing committees we need periodic face-to-face. I really believe that. It is the nature of this work. It is the consensus building. It just won't happen over the phone. Much less engagement on the phone and we need to be engaged here to reach that consensus.

And on the phone I don't know if anybody else is guilty of this but I know I multitask. And if you are multitasking do this stuff -- yes. No, I mean, you can't do that. Everyone understands.

DR. TSANG: It is harder to do it here.

DR. GELZER: Exactly. And the only other thing I would say this is a -- as I understand it, this is a public-private partnership. And the private piece of it, I mean there is a dues structure here. We pay dues.

So I understand the issues of the

government. I understand that we have sequestration but I honestly believe we are not a pure governmental agency and are not necessarily subject to those mandatory requirements such as AHRQ or one of the other agencies.

MR. SHEA: Thank you. If I could just comment on that. The situation is very fluid and we are hoping to get a bit more stable situation soon, predictable and so forth, whatever the final outcome is or the edict on meetings.

Over the past six months, we found ourselves in a situation where we either had to decide to pay for in-person meetings or we thought really risk the quality of the process or the further participation of people. And so we have gone ahead and done that. It is obviously not a great business model unless we find another stream of income to pay for that. And we kind of think like well if it is important enough to have a big process and

expert input, it is important enough to do it right.

But we are not going to -- I don't think anybody, whether it was your or us would stand for a compromise process. It just would fall apart. So that is not going to happen. And so we are thinking about all the options, including finding other money to do the meetings, even if the feds won't pay for it.

MR. WOLFSON: So I think about this a lot with my organization. So I think you have to put out olive branches to the government. One thing I would do with this meeting, I would not do a two-day overnight. I would make it start at one o'clock, go to nine o'clock, come in the next morning, just make it one overnight. You are reducing your costs.

MS. EAMES-HUFF: Those of us from the West Coast can't --

MR. WOLFSON: I just said one o'clock.

MS. EAMES-HUFF: We can't get here by one o'clock.

MR. WOLFSON: Two o'clock. Three o'clock.

MS. EAMES-HUFF: There is the time change.

MR. WOLFSON: Don't have people from California on this.

(Laughter.)

MR. WOLFSON: I don't know the regulations but some of the people around this table could have their organizations sponsor them and not charge the government to be here. And so I would --

DR. NELSON: Have the meeting in Las Vegas.

(Laughter.)

MR. WOLFSON: Have the meeting in Las Vegas. But I think if you ask some people around this table if their organization would pay for them, they would say yes, no problem. We want to be here so much that we would pay.

And to some of these organizations, it is affordable. To other organizations it is not and it would be a barrier and you wouldn't want to do that. But people who can pay, pay. And I think that is another olive branch to the U.S. government.

DR. LATTS: So this is the third panel that I have sat on and the issues that we have discussed have been incredibly complex. And you can't, frankly, follow the complexity over the phone. You know maybe sort of with modern teleconferencing capacity you could do it. But to be in those kind of facilities, frankly is probably as expensive as flying everybody out here for a meeting. It is very expensive to access the most modern teleconferencing capacities.

If you are reduced to having to do it over the phone, I would say this committee can't be this big. It has to be a third the size because you just can't process this large of a group on the phone, which then I think

you lose a lot of the nuances and the complexity of the discussion. So I think that would be a real shame but you just can't work a group this size over the phone.

DR. NEEDLEMAN: Lisa actually hit my point. I endorse everything that has been said. Phone is extraordinarily hard except for well-structured conversations or very short conversations around clear decisions.

But look around this room. You have got 30 people around this table, all of whom have made contributions. All of whom have made contributions. And that is not possible on the phone. I don't even think it is possible with videoconferencing.

So one of the clear tradeoffs in terms of an effective process if you are going virtual is you have got to go smaller and then you have to decide how much you lose from having a group of ten or 15, rather than a group of 30 involved in the conversation.

MS. YANAGIHARA: I just wanted to

maybe give some specific ideas about when you could do a phone conference and when is not.

I find when it is more informational, a little bit more unidirectional or just trying to clarify, ask questions, those kind of things, like Cheryl's idea about maybe we could have gotten together with the measure developers and kind of gone through the measure and asked questions and made sure we really understood it, that works pretty well on the phone. People are willing to ask the questions and things like that. But when you are really trying to have a conversation and really delve into the nuances of what this measure might mean, I think it is very difficult.

And we have tried videoconferencing. We do this a lot in California with our committees. Everything is multi-stakeholder what we do. And it just doesn't work. And having some people on the phone and some people in person, it just

doesn't work very well. And people will say that. They are like oh, I have got to be at the next meeting because this just didn't work at all. I couldn't really contribute.

And I think the minority view and those things that are really important to bring forward do get lost on the phone.

DR. WEINTRAUB: I want to contrast this meeting to a study section. Because I think what goes on here is actually considerably more complicated than study section. So if the NIH is forced to go that way, I hope we are not.

The American Heart Association study section is already by telephone. I am not on study section anymore but I understood from my colleagues it works reasonably well, with some problems.

Where it works well at the NIH is with special emphasis panels, which tend to be smaller. And those are already often on the phone. And I have been on those and it goes

okay.

Full study section, I think, when you have about this many people are even more sitting around the table. I think that it does impede the conversation but it is possible. But here, the complexity of the conversation is so extraordinary and the minority views need to be stated, as already has been said. But this would be more difficult than an NIH study section.

DR. GIFFORD: Echoing the complexity issue, I mean I chair a Board where in our organization we have nine different committees. And we do a lot by teleconference and video conference by phone but we also have to do face-to-face, and that is even where people know each other. And here where we don't -- even if you had the standing committee, you need to have some of that face-to-face. And I would echo the complexity.

When it is simple, when it is sort of straight forward, you can read something

and make a straight up and down vote with some discussion, you can do that by phone. Working through this is very, very, very difficult to do that without that.

And I wouldn't use the bathroom example because Nancy and I wouldn't be able to have conversations then.

(Laughter.)

DR. GIFFORD: It is not the body language. It is the side-to-side conversations. If you look around, everyone is having little sidebar conversations. And so I am leaning over and I am saying did we do that or where are we going. So I am trying to figure out where it is. It is really complex and you lose that.

When we are deciding what agenda items to have on our committees or other stuff, and even all the five other TAGs that I have been on here at NQF, you always say well that item we have to reserve for face-to-face. We sort of it is a little bit like

pornography. You know it but it is sort of like you can't have that on a call. You just need a face-to-face for robust discussion.

And so I think that would be the feedback I would give to CMS. I like Dan's suggestion you have got to give them something to go with but there are ways around it. And I think even when we are traveling, we are willing to work extra hours or do different times and to do that.

DR. LATTS: I wouldn't use the pornography analogy.

MR. SHEA: The Cost and Resource Use Panel said, you know it when you see it.

Larry?

MR. BECKER: So I go to a bunch of questions. And that is so what problem were we trying to solve? Are we trying to solve the Las Vegas problem? Right? And so you put very strict requirements on what we do and how we do it and this is sort of the way that is? Are we trying to solve the absolute cost

issue? And so are there trade-offs that we make in other places to take costs down because cost is the issue?

And then on the other side, what are trying to accomplish and what is the best way to accomplish that? And so in deciding how we proceed, I think let's take all of those elements into consideration and I think that it is at least a meeting type by meeting type evaluation if it isn't a meeting by meeting evaluation.

DR. MARCINIAK: So I think the only thing I would add is when you look at the course of the meeting we have had over the last almost day and a half now, the question also becomes one of meeting efficiency.

So I have been no the study section type calls. I have been at meetings like this and I have been in meetings at the corporate level. And the question becomes the discipline around the meeting itself.

You know we said the same thing a

lot of times over, and over, and over again. And there is nothing particularly wrong with that but the challenge becomes well how do you compress that all if part of the problem is well we are bringing people from the West Coast, we need to efficiently use their time. We have people who come to Washington because after NQF we have three or four other meetings we are going to. So there is also a meeting efficiency issue in terms of how we elect as a group to sort of drive ourselves to consensus.

MR. SHEA: Thank you, very much. This was really much more than I had hoped for but it was exactly what I wanted. This is very helpful for us. We are going to be as responsive as we can be to the people who are putting up the money -- it is our money after all -- to do this. We aren't going to compromise the quality of the program. Obviously, the difficult thing is well sort of what can you do here. What can you do this

other way. I kind of think we are going to wind up both using some more virtual technology and also finding some other sources of money to make sure we are able, when we need to do a meeting, to do a meeting.

But this is extremely helpful to us. We have, as with all the meetings, we have a record of this. We are going to draw up a summary and we will be using it internally and also some of the points when we meet with the HHS folks. So thank you very much.

MS. TIGHE: Okay, great. I am going to jump right into the harmonization discussion. And while I am going over some principles for the Steering Committee to review, if the developers from CMS and Health Partners want to get in position with microphones --

MR. AMIN: One thing. Just as we introduced here, clearly the harmonization discussion is intended for measures that are

recommended for endorsement. The measure that we are having a conversation about the committee felt very uncomfortable about making a final decision. You will, obviously, have a time after comment to discuss whether or not you want to, after the comments have been provided by the membership, to continue to move the measure forward. You know, again, thinking about today's vote as sort of preliminary until the third stage, which will be after the comment period, which it will be your final vote to the CSAC and then it will go to CSAC.

So the purpose of today's discussion is really intended to be, at this point, sort of preliminary, depending on what the ultimate decision of this committee is after comment. So we, obviously, don't expect it to go as detailed as it would have if the measure was endorsed but we don't want to miss this opportunity to miss this opportunity with both the measure developers in the room and

you guys just having thought about this measure.

The Medicare Spending Per Beneficiary measure is not going away. And so we want to make sure that we continue to push forward our agenda around public-private alignment. And so that is the nature and the reason why we are continuing this discussion today.

I will send it back to you, Lindsey.

MS. TIGHE: Okay. And on that side as part of the documentation of this discussion, you have a printout of a side-by-side table to compare the two measures we are discussing. Well you got it yesterday. If you don't know, let us know and we can provide some additional copies. It was printed out. It was also in the packet that we sent out.

So I will just jump in. Our process for this conversation is that we are going to just review the principles for

harmonization and give an overview of what related measures are. We will ask the measure developers, they provided a joint response letter to you all which you received your pre-meeting materials. We will just ask them to provide a brief presentation on the rationale for the measures. We will as Staff walk through some of the conceptual and technical similarities and differences, we will open it up to committee discussion and ultimately committee recommendations for areas for harmonization or justification for not recommending harmonization.

Well I will just keep talking. So why do we harmonize measures? It is to identify the components of the measures, which can be standardized for consistent measurement of a population, condition or resource. We asked the developers to begin this effort prior to submission of measures to NQF. These developers have working together for the last few months to review possible areas for

harmonization between their two measures. And we are doing this for related measures that share similar characteristics and measure focus, such as the measure types, the same cost and resource use service categories and the same population.

This reiterates what I said on the last slide and then we are just looking to the Steering Committee to assess the value if the differences and specifications are necessary or unnecessary. And then the burden if the differences and the specifications infect the interpretability, if the differences affect the data collection burden for those who are being measured.

The desired outcomes from harmonization, we are looking for consistency in the measure results, interpretability across levels of analysis and data sources, reduced burden for the providers and implementers and improved interpretability for the patients who are looking at the data.

Principles, we look at the conceptual harmonization of the measure. So whether the measure intent, focus, or target population is the same or needs to be harmonized. And then we also look at technical, which is how that measure, intent, or focus is operationalized through the specifications, data elements, code sets, things of that nature.

We don't look to harmonize the statistical risk adjustment, the risk stratification or the statistical methods for estimating measure results as of our current guidance.

And if there is no harmonization recommended, then we ask that the committee just elaborate on the value of the different concepts and different technical specifications, making sure that the benefits outweigh the potential burden or risks of having un-harmonized measures.

And that is it. We will turn it

over to the developers, HealthPartners, and CMS to just provide a rationale for the two measures.

MR. BALLOU: So believe it or not, I am still here. And I would like to introduce basically the key harmonization issues that we have talked about with our colleagues at HealthPartners. The measure names are somewhat long. What I am going to refer to as the CMS measure is the one that we just discussed, the Payment Standardized Total per Capita Cost Measure for Medicare Fee-For Service Beneficiaries. I am going to refer to the HealthPartners measure, that is the Total Resource Use Population-Based Per Member Per Month Index. That is NQF Measure 1598.

So those are the two measures that we have gotten together over the past month to two months to try and identify key differences in the measures and discuss which among those differences might be amenable to harmonization.

So I would like to quickly take you through some of the key differences. And then while I am summarizing this, what I am doing here is I am summarizing the memorandum which I believe, in addition to that table, was included in the 400 plus page packet that was emailed to you. I am doing that summary but obviously, my colleagues at health partners can speak for themselves. So after I am done with that, I would like them to have an opportunity to respond.

The key differences that we have identified are essentially five in these measures. They are both total per capita resources use measures. Total comes with an important qualifier that I will come to in a moment.

First of all, they focus on different target populations. I will go into a bit of detail on each of these but I am just listing them for the moment. They use different risk adjustment and payment

standardization methodologies. They attribute patients differently and they differ in their inclusion of pharmacy data.

The upshot, and again we will get into a little bit more detail and then the discussion will ensue, but the upshot of the memo is that we, meaning CMS together with health partners have jointly recommended maintaining each of these differences in the respective measures, rather than harmonizing.

And the reason for this is that really the key difference here is the target populations. The differences in the target population drive a lot of the other differences that we see. The target populations differ meaningfully. The CMS measure, as you know, is designed and tested for Medicare fee-for-service beneficiaries. It is a very specific population. It is an older population in general and those who are enrolled in Medicare who are not over 65 are not representative of the population under 65.

Many of these beneficiaries have multiple chronic conditions. They tend to have heavier use of inpatient hospital and post-acute care services than commercially insured populations.

The HealthPartners measure, by contrast, is designed for, tested on, and endorsed for commercially insured patients. And so there is a different population there. Again, younger and generally healthier. You are looking at somewhat different mix of services. You will, obviously, much more maternity and newborn related services in a commercially insured population than you would see in the Medicare population.

Because these target populations differ meaningfully, components of the methodology have been tailored to make the resource use measurement for the respective populations as accurate as possible in two respects. First of all, risk adjustment. We use different approaches. CMS uses the CMS-

HCC risk-adjustment methodology. And as the discussion goes on, if needed we can get into reasons why. HealthPartners, on the other hand, uses Johns Hopkins ACG risk-adjustment methodology.

So different approaches that are intended for different populations; Medicare versus commercially insured.

Similarly, the payment standardization approaches are different. So in the NQF terminology these are not per capita cost measures. They are per capita resource use measures, which means that you are somehow standardizing services to make comparability across different types of services. And you need a payment standardization algorithm to do that. CMS uses its agency-wide methodology for fee-for-service beneficiaries. It is extremely specific to all of the prices that are paid for really all of the services that Medicare covers.

HealthPartners uses a total care, their own Total Care Relative Resource Values methodology, which is very appropriate for the commercial population on which their measure has been tested and again is likely to get the relative prices right in that population, much better than say the CMS approach would be if you tried to take Medicare standardization and apply it to a different population.

So our view jointly is that if we harmonized on either risk adjustment, which we realize per Lindsey's observation of a couple of moments ago is not necessarily recommended, or is not recommended for harmonization or payment standardization, this would reduce the accuracy of one measure or the other. So we advocate maintaining these distinct approaches to risk adjustment and payment standardization for the two respective measures.

And the other differences I am not sure I need to cover in as much detail at least by way of overview because they have

been covered so extensively in the earlier discussions. The first is that we have different patient attribution rules. We have discussed extensively the attribution rule for the CMS measure. The HealthPartners approach is also a primary care-based rule but it is somewhat different. When they were up for endorsement, they were given the option of listing the attribution approach as a guideline or as a specification and the measure valuation form has changed since then. CMS did not have that option in the most recent form. So we listed the attribution rule as a specification. But this is a rule that, as we discussed, is being adopted across other agency initiatives. So CMS has an obvious programmatic interest in maintaining possibly in an improved form the rule that is currently before us.

And then finally the CMS measure does not include pharmacy data for reasons that were discussed late yesterday. Many Part

A and B beneficiaries are not enrolled in Part D. Whereas, the HealthPartners measure does. I think we are all in agreement that if you have pharmacy data for your patients, you want to use it. And so in our view it is appropriate for HealthPartners to continue to include it but that it should not be included in the CMS measure.

So that was the basis for our conclusion that we recommend against harmonizing on these four or five dimensions. But again, I would like to give HealthPartners and opportunity to also weigh in.

MS. KNUDSON: Thanks for that summary, Jeff. Hello, everyone. I am Sue Knudson with HealthPartners. I lead Health Informatics there. I am joined by my colleague to my right, Chad Heim also from Informatics and Gary Kitching to my left.

I would just like to make four additional points. We worked together on the summary that Jeff had made.

The first point is we all know who CMS as providers and healthcare professionals. You may not be familiar with who HealthPartners is. So I just wanted to briefly tell you about us.

We are a consumer-governed non-for-profit organization based out of Minnesota. We are an integrated finance and care delivery organization. And why that is important to the relevance of our use of these measures is because we are not just simply reporting to others. We also need to make sure the measures and information are usable for our own practice in hospitals as well.

The other thing that I would just like to point for many of you who aren't as close to the Minnesota market, because we own and operate both care and financing, neither organization is exclusive to one another. Our care delivery footprint is a multi-payer footprint that is not a closed model, HMO, staff model traditionally. We are practicing

in the open market, as is everyone else in Minnesota. And likewise our payer is also not exclusive to the care delivery system. As many of our partners here from Minnesota can attest to, we partner with everyone in the region.

So with that, Jeff had also mentioned the second point I wanted to make was just the aha that the Steering Committee came to yesterday with regard to the CMS measure really being calibrated to the payments. And so that is inherently a different resource use methodology than we have employed to build our tool, Total Care Relative Resource use, which is really a relative system across the full continuum of care, folks seen simply on resource use. It is a patented methodology that we spent over a decade developing. We have it out in the public domain now free of charge. So for others to use.

The third point I wanted to make

was Jeff had highlighted that yes, our measure does include the pharmacy data as it relates to your interpretation of the grid that was handed out. What is different about the commercial population and this measure is we are not excluding any information. We are including -- it is a very patient-centered measure. It includes all care for taking care of patients and members that we have that administrative claim data for.

What is very different is the prevalence of types of services like long-term care, skilled nursing care, which is much more prevalent in the Medicare population. It has some prevalence, not as much, obviously, in commercial. So the point I just wanted to emphasize, no exclusions in that regard. That the differences based on the care needs and the use patterns are much different.

And then lastly, Jeff also pointed out with regard to attribution, the CMS measure went through with that as a

specification. Ours, indeed, was offered as a guideline. We shared our methodology in the spirit of transparency but also in the consumer and commercial market, understanding that different areas around the country may have rules that makes sense in those markets.

And just by way of example there, the work that we have done in Minnesota since the endorsement of the HealthPartners measure in 2012 early in the year, we have worked with our community collaborative to review all the attributions and have landed for community reporting on a single method because, essentially, what we found is you know they returned roughly about the same result. So it was an opportunity for us to standardize that.

So again, I just wanted to emphasize that the attribution method for the HealthPartners currently endorsed measure was a guideline.

So with that, I think we would entertain questions.

MS. TIGHE: So just on the next slide, Staff briefly went through and just highlighted some of the areas where these measures have overlap.

DR. RYAN: I have a question for HealthPartners. Have any other commercial payers adopted this method and can you speak to their experience using it?

MS. KNUDSON: Yes, thanks for that question. Right now this is the resource use methodology that we have endorsed. We have at least, I think we are up to about 60 users across 21 states.

DR. GARRETT: So I have a question about continuous enrollment. So Sue, my understanding is in the HealthPartners measure there is a nine month continuous enrollment requirement. Is that right?

MS. KNUDSON: Yes, Nancy, that is right. And that is with regard to accumulating enough diagnoses in the commercial population to make sure we have the

proper amount of data to do good risk adjustment.

DR. GARRETT: Okay. And so for the CMS measure, is there an equivalent?

MR. BALLOU: Well again, for the CMS measure, we are at a looking at a full year.

DR. GARRETT: And the beneficiaries have to be enrolled for that whole 12-month period continuously?

MR. BALLOU: They need to be continuously enrolled in both Parts A and B, correct, for the full 12 months.

DR. GARRETT: Okay, thanks.

MS. TIGHE: So I'm going to jump in. I just wanted to highlight what Staff has laid out as kind of some of the similarities between the two measures and differences. The target populations do have some overlapping in the age ranges, given the commercial population, the Medicare population aren't strictly defined by age. There are similar

settings. They are non-condition-specific and they both look at a one-year time frame.

The measure focus, they both use standardized costing approaches, although they detailed the differences there. They both are per capita measures. They both hit on similar resource use service categories.

The technical, the target population, we are looking at a Medicare population versus the commercial populations and the measure focus. There are some differences in the costing approach as they have outlined and then there are also a few service category differences that are related to how the measures were specified.

So just moving to the next page, we just really want the committee to take the time to consider whether or not these measures should be harmonized first and then from there what those recommendations may be.

So to focus in on do the measures have sufficiently different populations to

justify two measures. Are the standardized costing approaches different enough to justify having two measures? The justifications and rationale provided by the developers, is that sufficient for justifying two measures? Are there areas where maybe these measures should be harmonized? And then the last question is perhaps a little bit of moot point, at this point.

MS. CLARK: Just a question on the harmonization process. So are we when we are talking about that, are we saying we have an option of keeping two completely separate measures or we can take each of these components and say some components should be the same? Is that what we are saying?

MS. TIGHE: Yes, so you have the option of keeping two measures as they are, based on the justifications from the developers. You could recommend areas for them to harmonize related to the measure conceptual focus or technical specifications.

Or an extreme version is to pick a best in class.

MS. CLARK: So I guess just another comment then on the approach to the costing methodology. I mean, those are completely different. So I mean I don't see how we can say that those are equivalent. I just had a question about that. I mean, as I understand it, you are using relative value units where there are some that exist and where there aren't any that exist, you are using billed charges, relative billed charges. Is that right?

MR. KITCHING: Yes, thanks. Actually, we impute those values based on the billed charges. So you are correct but actually it is just to create the actual relatively is between the service codes, yes.

MS. CLARK: So that is how it is done for like the pharmaceutical costs?

MR. KITCHING: I terms of creating the relativity as across but then we actually

go back to get the paid, we blend them differently again, just to make sure we get back. We want to get back to paid amounts at some point. So we do that.

MS. CLARK: Okay, so the final measure is in dollars or RVs?

MR. KITCHING: I reflect it as RV use. This is an RVUPMPM.

MS. CLARK: Okay. So that is another distinction. It is a per member, per month and not an annual per.

MR. KITCHING: Well I would say you could characterize it as a PMPM but you can reflect it as a per member per year. I mean, it is just how you reflect it.

MS. CLARK: Okay.

MR. KITCHING: So, yes.

MS. TIGHE: Larry, Jennifer, and then Andy.

MR. BECKER: So two questions and maybe -- well, let me just ask the questions. So the first question is so what do the two

developers think about harmonizing? And secondly, has anybody looked at what the practical effect is of trying to put these two things together and what comes out the other end?

MS. TIGHE: I will let the developers hit on that first.

MS. KNUDSON: I will comment first. This is HealthPartners.

If I put care delivery taken action off of this hat on, I would prefer measures calibrated in tune to those populations. So when I really think about usability and taking action on it to make improvements, I think about having the measures tuned, if you will, or calibrated to my commercial environment, my state public programs, if I have a method there, which we are doing some development work to tune our method to our state public programs, as well as look at the Medicare population separately. So that would be one response to that.

The second piece I would add is when HealthPartners went through this process a couple of years ago, many of you who were on that Steering Committee will recall that we have two endorsed measures. One is a total cost of care measure, which is actually based on allowed payments, member liability, as well as planned liability. So the power in terms of taking action on these measures is best when they are used together because we find the actionability with understanding total costs and then we work with practices, including our own, to understand that total cost performance and all the things that have been debated here over the last couple of days.

But then this resource use measure really helps to benchmark practice opportunities in a way that is not confounded by price differences. So we really see the power in using the two measures together and that is how we have actively used them.

So I hope that helps.

MR. BALLOU: Nothing substantive to add, other than to reiterate our view from again the Medicare Population perspective that having a highly customized, highly tuned standardization approach is important for accurately reflecting relative resources.

DR. STEPHANSKY: So what would be the result if we said yes, you have to harmonize these two measures, what would you do?

MS. KNUDSON: Well you know the first thing I guess is -- you know, I don't know that I can, in fairness, answer that. If we start with the risk adjustment question, for example, you know the NQF guidelines, I think as we looked at those around harmonization, too, it was a natural response to us that they are not conducive to be harmonized. The risk adjustment model used by CMS is fundamentally different than what we are using for a commercial population. The

disease prevalence is difference. There is not an open source risk adjuster that would work in a way that would have the results be to the level of accuracy that we need them.

We have been open to testing our measure, using an open source risk adjuster when it is available in tune to the commercial population but as of yet, that is not available. So that is one way to respond.

You know, so I technically don't know if it is possible.

MR. BALLOU: Right. And again, I think I would echo that response. And even if it were mechanically doable, you would get numbers that, in our view, would not make sense.

MS. EAMES-HUFF: So I will just say from the get-go, I think there probably are enough differences between these measures that it would be difficult to harmonize. I have heard though from quite a few people that there is a desire to look at the commercial

population resource use and what is happening in the Medicare population resource use. So I am interested in you commenting on is there some level of comparability or are they so different that you can't really compare the numbers?

MR. BALLOU: Are you asking about comparability of results --

MS. EAMES-HUFF: Yes.

MR. BALLOU: -- even on a non-harmonized basis?

MS. EAMES-HUFF: Yes.

MR. BALLOU: I guess I would argue that I would go back to our initial point that the differences in the target populations are sufficient that one should not make other things equal source of assumptions in making those comparisons.

I'm not sure if that -- it sounds like that doesn't quite address that.

MS. EAMES-HUFF: It doesn't quite satisfy because I think we see we do with

quality measures, we compare the commercial population to the public population.

So there may be reasons why it would be different, you know, that drive different cost behaviors. So we would see a difference. And I think it is just understanding that and whether or not these two measures, looking at them, could shed light on that or not. I'm not sure if that is possible.

MR. BALLOU: Yes, I guess my take and HealthPartners might or might not agree is that to do that effectively you would need what again we don't see being available yet, which is a risk adjustment algorithm that can accurately and adequately capture the full spectrum of the populations that we are talking about and similarly a payment standardization methodology that can do that.

So these measures are very distinct in their approaches right now. There isn't one approach on either the payment

standardization or risk adjustment side that would allow us to essentially throw everything into the same comparison group. Once you had more valid methodology for either and risk adjustment and payment standardization, then you could certainly do valid breakouts of differences between the Medicare experience and the commercially insured experience.

MS. TIGHE: Andy, Mary Ann, Nancy, Cheryl.

DR. RYAN: So I think for payment adjustment and risk adjustment -- payment standardization and risk adjustment, it makes sense to not harmonize the measures. It makes basically zero sense to try to harmonize.

The one place that it seems like there is a reason to do it is that they are attribution because it is complicated. We have talked at length about it on this committee and the idea of providers having different rules for different populations, it seems like that would be extremely difficult

to manage and would create some kind of, I think, challenges, and kind of comparing costs kind of apples to apples based on their accountability for Medicare and non-Medicare patients.

And I also just kind of wonder from HealthPartners perspective if they think that their measure might just kind of become irrelevant. If Medicare is moving along with this other way, it might be just better to kind of get in line behind what they are doing. So that is kind of my question is around attribution and whether -- how that will play out for providers who are being profiled.

MS. KNUDSON: So in our experience, and we come from a market that has a rich history of measurement and transparency around it and collaboration, even though it is a very competitive market as well, although we tend to compete around the things that matter and sometimes these definitions are not in

that domain.

So for attribution what we use is attribution and what we have recently come to is looking at the visits and it is looking at most visits. It is primary care, as Jeff had mentioned. We do include the nurse practitioners and others that were discussed.

And then more recently in our community collaborative discussions, we have agreed to a look-back period so as we are all creating new and more improved care designs that we also don't incentivize requiring a patient to come in on an annual basis, if it is not necessary, moving and taking advantage of virtual care and other online services that can be done more cost effectively. So we are trying, over time, adapt our attribution models to make sure that they are moving with the care designs. And so to that regard, with all the reform going on and the ACO work and everyone really working to improve and hit on all cylinders with the triple aim, this has

got to move with it.

And so those are just some of the recent innovations that we have done. We, too, as Brent Asplin had spoke to earlier, are familiar with this CMS attribution. Our colleagues in our care group from Park Nicollet are also a pioneer ACO. So we are familiar with those different levels.

Then the last comment I would make about commercial is when you are looking for adoption across the country, the commercial market is different. The payers are different. The product design is different. There might be different longevity in certain plans. So it is not as though I am 65 and I am now enrolled in Medicare and that is static across the rest of my life. There is movement in the commercial market. And so we just need to be mindful of those things in talking about one standard approach.

MS. CLARK: I am just curious. So HealthPartners, I assume you have a Medicare

Advantage plan. Is that right or no?

MS. KNUDSON: As a payer, we historically have not had a Medicare Advantage plan. We are a cost plan. So as of right now, that is why we had not used this measure on Medicare because we don't have the whole suite of data within our operation to run it.

MS. CLARK: Okay. I was just going to ask you what approach you would use if you did have. Or CMS, what are you envisioning, you know Medicare Advantage plans, how they would adopt a measure. I am assuming it would be the Medicare measure.

MS. TIGHE: Nancy?

DR. GARRETT: So a couple thoughts here. First of all, as a provider I think I kind of think of the cost measurement as like maybe ten years behind where we are in quality measurement right now. It feels a bit like the wild west where each commercial payer is taking a different approach to all of these measurement issues around attribution and how

to actually define these measures. So really thank you to HealthPartners for advancing this and starting to move towards a standard. I think it is really important work.

So specifically on the issues we are talking about harmonizing, I think with risk adjustment I agree that it really makes sense to have a different methodology for the different populations that that works in that situation. And so I don't think we should be trying to harmonize there.

With some of the other pieces, I think we need to have some more thoughtful discussion. So for example, I agree with Andrew, that is a place where, again, as a provider when each payer has a different approach to attribution that is really difficult. And so is there a way that we can start to move towards a national standard around attribution? I think that is an important goal. I don't think differences in the populations justify differences in that

approach for attribution.

With pharmacy, you know, not having data on pharmacy is a common problem for commercial payers also because a lot of employers carve pharmacy out. And so there are approaches where you can calculate the measure with people who have a pharmacy benefit and without and then create a blended PMPM result.

And so I think that is something that CMS should consider as well, so that we can have a measure that includes pharmacy. We talked about in the previous discussion that pharmacy is such an important component of total cost of care.

And then the last comment is around the different approach to the payments. You know, as Sue said, CMS is really calibrated to the payments and HealthPartners is calibrated to resource use. And some of that has to do with methodology and the difference in the way the reimbursement works

but some of it is a conceptual difference in the measure. And so given that conceptual difference, I worry that it is not going to be very clear to providers and even consumers as these things get rolled out what that difference is. So I am just not as convinced there that we don't need to consider harmonizing because of the different population there. It seems like there is a conceptual question. What are we trying to do with these measures and what is the reason for a difference in the approach?

So those are my thoughts.

MS. TIGHE: Cheryl, Jack, and then Brent.

DR. DAMBERG: I was wondering if the measure developers could comment on, because I agree sort of the underlying methods used to generate the measure are very different and probably should stand. But I guess I am wondering is there an opportunity sort of kind of downstream, once you generate

the end result, is there any way to combine these so that if CMS is giving feedback reports to individual physicians, the HealthPartners on the commercial side could be integrated such that there is some way to translate your metric onto the same metric, if you will.

So I don't know if you guys have considered that. Like is there any kind of common denominator in there to allow that combining of information?

MS. KNUDSON: I am not sure if I can answer that directly because we haven't specifically looked at that but if I could address kind of both comments together, Nancy, specifically your comments with an example.

The reason we have calibrated to resource use and see the power of using this in addition to the other measures that we have created is around taking action. So for example, we could be looking at the use of imaging in a population as a driver of overall

cost and we could isolate that the resource use really is predominantly in a hospital outpatient environment versus a freestanding. The way we have calibrated it based on resource use, the resource use for that scan, meaning it is the same scan in this example but the difference is price. Generally speaking, the hospital is a more expensive price for that scan, versus an outpatient ambulatory kind of setting.

And so by isolating resource use to specifically that, it helps to make place of service decisions that can improve overall cost performance. And so that is why that is sort of that actionability component was the rationale behind us using a pure measure around that.

So then if I go to your question, Cheryl, I would say to the extent that practice patterns in general hold firm in our care delivery environments from our commercial to our Medicare patients. If I always send my

patients to the hospital for a scan, I am probably not going to be changing that by payer. So to the extent that one source of information out of in this example, commercial could help you understand those practice patterns, it could have some carry over effect but we have not studied it specifically with the measures.

DR. DAMBERG: Yes, I am thinking more at this observed to expected. So if you are going to score me is greater than one or less than one, it is that consistent across. And then obviously, maybe the drill down reports look slightly different, so that they can tease apart where those differentials are coming from. But I was trying to figure out is there one standard metric where these two could be married in some way, without having to change up the underlying structure.

MR. KITCHING: I think our position in a lot of the work that we do, is you have to make sure that you separate the

populations and you keep them separate. So like the Medicare population is measured amongst the Medicare folks and the Medicaid is the same thing, and the commercial as well. We don't really see value in actually presenting a number that actually has all those three blended. But if you were to do it, you would blend it by population. But again, you have got to keep them separate when you measure them separately. So a commercial population is measured against commercial population and the Medicare against Medicaid. And you rank them as such and then you can blend those two based on populations underneath.

But again, as HealthPartners, we just don't see any value in actually doing that because a person isn't a Medicaid member and a commercial member together from a transparent perspective. They are different people. So actually when you are going to present that information, they should be

separate all the time. So that is a perspective that we take.

MR. BALLOU: And I think we would share that perspective. Again, in principle, we have two measures here. And as you know, when we discussed earlier in the feedback reports that we provide to physician groups, we provide many measures. An illogical way to potentially approach this if one were in an environment where one wanted to study a broader population than either of these measures currently cover, you would simply compute both measures, according to the measures respective specifications. And then you would report them with the appropriate corresponding drill down, I think.

But again, as this group knows that the focus of the CMS measure, at least at the moment is limited to the fee for service beneficiaries.

MS. TIGHE: Jack, Brent, and then tom.

DR. NEEDLEMAN: Okay. To cut to the bottom line, I think it is probably premature to try to harmonize these measures because we don't have a basis for choice among the differences in them. That said, I think some of the differences are probably smaller in practice than they are in principle. So I know that there is an extensive literature that looks at different risk adjusters and tries to compare their performance. And I suspect that when you look at the impact, the ultimate impact at the group level or the provider level once things are aggregated up, they don't provide that big a difference in the results. They are all using different nonlinear combinations of all the information that is available from the administrative data to try to get an estimate of likely use and they do it slightly differently and they produce slightly different rankings but not all that different.

But we don't have a basis for

choosing ACGs versus HCCs. That is so compelling that we ought to say do it one way versus the other. There was a lot of discussion in the first panel about the high cost of the ACG methodology to people and that was a real consideration. I would love to see an open source, low cost publicly use version of a risk adjuster that does well enough.

But we aren't there yet. And the issue there is doing enough work with large datasets to understand in practice what the differences are in risk adjusters in terms of the rankings, in terms of where providers are classified, where patients are classified, and what difference it makes in practice when these measures are used. And we haven't had that.

So I would encourage both sets of developers to actually start looking and start reporting what difference it makes to use one risk adjuster versus another. One standardization, one standardized price

schedule or resource schedule versus another.

Again, the last time last year we had an extensive discussion about standardized prices versus non-standardized prices for whether it is useful. It could be comparing -- I think the comparison was Memphis to Peoria. And don't ask me why we wound up with Memphis and Peoria but there we were. Memphis I remember. Peoria I may be making up. Was it Minneapolis? Memphis and Minneapolis. Thank you, Helen. But the same thing.

So the real thing we have discussed about standardization is the compression of differences in actual costs and resources used by the providers in settings and how that is masked by different standardizations. And that was alluded to in the sense of if you use the same standard price for hospital-based imaging or lab tests and freestanding imaging or lab tests, you are losing a source of variation in what is being paid and what costs are being realized in the

system.

So standardization does that. We have to understand where standardization is beneficial in terms of understanding resource use differences and where standardization is masking important resource use differences. We aren't there yet in terms of recommending a single approach to standardization. So again, I would encourage those who are using these measures to use alternative ways of pricing them to see what we learn and don't learn from them and start reporting that so we have a better basis for making choices in the future, rather than simply looking at the choices you have made and trying to infer what is gained and lost by them.

With respect to the drugs, I am going to second Dolores here. I think the lack of drug information with your Medicare beneficiaries is a significant weakness of the measure you have got. And I know because Part D with different companies makes it very

complex and perhaps more complex than the simple carve-outs with the PB, Pharmacy Benefit Managers. In the commercial market, it is harder to do.

But we have seen some approaches. Again, last year the Ingenix folks said we stratify. When we have got the pharmacy data, we report plans with pharmacy data. When we don't have pharmacy data we report it without and it is two separate sets of measures. And you have got 60 percent of the folks in Part D. Figure out how to get that data reclaimed into a total cost measure for the beneficiaries and stratify it as a first cut version of how to do that. And I think that will move things along and you will have two or three or four years of challenge in figuring out how to actually get useful data from the Part D folks who don't want to give it up. But it is an important thing to do and that needs to start.

So as I said when I started, I

think right now harmonization is premature but there are lots of explicit things that we ought to be encouraging the measure developers and others to do to help us to understand what the differences are in these measures and what difference it makes in practice so that down the line we can think about what preferred methodology should be adopted.

DR. ASPLIN: What he said. I think Jack summarized quite a few of the points was going to make. I accept the arguments of the developers that we don't need to harmonize today. And along the lines of what Jack was saying, I am probably more interested in what is going to trigger the work that needs to happen over the next five to ten years so that we would have the information in front of us to perhaps make a different decision in the future than we are making today on this harmonization point. We need to learn.

And I will say that HealthPartners

does a great job of having the very conversation you just alluded to today with the Liberty Systems. So when we sit down, we have not only the Resource Use Index data around resources but there is also the Total Cost Index data. In fact on Monday, I had a conversations with one of their medical directors about both of those components. So they look at -- so we understand both how our pricing decisions affect total cost of care, as well as our resource use decisions.

So they are having those conversations, even though the measure itself uses the resource standardization process that they got through NQF. So they do a very good job of that. And I would hope that we are going to get the information we need to reevaluate the decision down the road.

MS. TIGHE: I am actually going to jump in really quickly just in the interest of time. It sounds like we are kind of coalescing in the idea that these measures

don't need to be harmonized today and that there is a need for more information the future. If there are any points that need to be made in addition to that, let's go ahead and make them but we do have the whole Day 2 agenda to tackle.

MS. YANAGIHARA: Very briefly, I think that there was a question about whether there is value in being able to compare across product lines and I just want to iterate what I hear from the physician groups in California there would be value. So if it is possible to get to something that is comparable and combinable, they think of their whole patient population. They don't think in terms of which product line. And so there definitely would be value. And I think the question is can we get to a measure that would be comparable across product lines.

So I just want to just keep it in mind that it is an important need and so we shouldn't give it up.

DR. TSANG: Just two quick questions for HealthPartners. Understanding that HealthPartners probably has very, very little leakage, how that actually impacts on some of the attribution issue. And then secondly if you can just expand a little bit about what you have learned from implementing this measure and have you used it and how you would actually -- I am sure you had talked about how to improve upon it as well.

MS. KNUDSON: So I assume you are referring to our own delivery system in terms of leakage. Actually, we have an interesting delivery system because our ambulatory clinics, the Legacy HealthPartners Clinics, if understand the Twin Cities you know we are very divided by rivers and our clinic footprint is much on the east side. But we do have some clinics no the west side as well. Our hospitals are all on the east side, with the legacy organization. And so that means we partner with others in the community for

specialty coverage as well as hospitalizations that are not in our own care delivery, our own delivery system.

So but for our own delivery system, as well as those that we contract with through our health plan, a part of the suite of our information is to help them understand who their referral partners are and what their Triple Aim performance is as well. And so it is a very transparency-enabled discussion again.

We do have a fair amount of keepage, if you will, to use the term more positively but we do have a rich history of partnering with other specialists and hospitals in the community as well. We just track that very closely.

And then in terms of the lessons learned question, what we have very early learned is this is a dialogue. It is not just us going out and calling the score with those that we partner with in the community. It is

really sharing the information, having a dialogue about it, having an engaging discussion about how the measures can be improved and evolved, understanding practices, all with this evolution of really keeping the patient at the center. And once you do that, there is less chance for things to go awry.

And we have also learned you know at the high level, the precision and the accuracy is very good when we start drilling this stuff down to really hone into opportunity areas. You know multiple payers have different versions and what we are hearing from providers is yes, the HealthPartners report might be different from my Blues report but they are directionally the same. And it is those providers who are taking action on those directional consistencies that we are seeing have the best uptake in terms of improved performance.

And then the last comment I would make in terms of lessons learned is we very

intentionally pair this information across our health plan applications with quality and patient experience information. So it is truly Triple Aim reported. And that for sort of gaining the hearts and minds particularly of our clinical teams, that has resonated very well.

Does that help?

MS. TIGHE: Okay, great. I'm going to use that to end the harmonization discussion now and move on to our discussion of risk adjusters.

So this will be a conversation with Taroon Amin and Karen Pace who are both NQF Staff leading, and then we also have Syed Mehmud joining us. You can come up to the table.

And then on the phone, Operator, I believe we have got Steve Frank from Optum, Greg Pope from RTI, Chris Tompkins from Brandeis, and David Bodycombe from ACG.

MR. AMIN: So while everybody is

joining us onto the table, I just want to set a little bit of the context of this conversation.

So as we know, this is sort of newer area of measurement in terms of national consensus standards. The first iteration of the Steering Committee provided a number of -- and so we are moving into much more of a conceptual discussion around cost and resource use measurement in general. And so we are using this time with the Steering Committee to really think through some of the conceptual assumptions and recommendations that were given from the first Steering Committee.

So as background, as we first initiated this work noting that we have the refresh of this group and we have new experts joining and others that have stepped away. So it is a good time to rethink some of our assumptions as we worked on this effort from the beginning.

The first cost and resource use

effort put together a series of guidance and also the measure submission form, which included various different components. Those components allowed guidelines and specifications, which some of our colleagues, our developer colleagues referenced earlier.

But specifically there has been a number of questions that have been raised in the field related to our current guidance related to risk adjustment. And this guidance is consistent across quality and resource use and we want to kind of bring up the issues that have been raised in terms of our guidance as it relates to the need in the community of various different stakeholders.

So we have invited a number of experts who are developers across the spectrum, developers who have participated in our first effort. Developers who were working on the Medicare grouper, additional experts who have participated in earlier conversations to provide additional context. And so this

will be a little bit of, obviously, committee discussion but we would at various points ask our colleagues that are on the phone and in the room to provide their perspectives on what they are seeing from the field.

And I am joined here also from -- Syed has joined us from the Society of Actuaries. Thank you very much for joining us. The report that he co-authored, the 2007 Risk Assessment Report is often cited in the specific area of looking at the relationship of various different commercial groupers that are available in the field.

And I am also joined by our lead methodologist, Karen Pace, who has worked in various different components of cost -- well cost most recently, but also in our quality side.

So the guidance that we are looking for in the Steering Committee is to reflect on these issues first in terms of our guidance, the input that we are getting from

the community related to our guidance, and then to advise the CSAC who is ultimately responsible for setting our criteria and guidelines in our measure submission form.

So with that, I am just going to get started and we can ask kind of questions as we go. I know that was a lot of preamble.

But as we discussed, various different HealthPartners measures, and I am using HealthPartners really as more of a case study. We are not having a discussion about this measure in particular. And I should also preface this by saying NQF in a lot of ways is very appreciative of the leadership that HealthPartners has taken in this field of cost and resource use measurement first by participating in the first cost and resource use effort but also in really working with various different communities across the country to try to implement this measure across various different sectors. So they have obviously, shown a lot of leadership in

this space and we want to be responsive to the needs in the community related to cost and resource use measurement.

So as background, the HealthPartners measure uses the Johns Hopkins ACG Risk Adjustment approach, which we discussed prior, which is a commercially available risk adjustment model. And so I think that is pretty clear. And we can go into questions as we talk about this.

And various community collaboratives, providers, consultants to health plans have worked with the developer in order to use various different risk adjustment models that are standard in their community. And so various different, we have heard this as well from various different risk-adjustment methods folks. So in some communities they use the ETG. In others they use the DxCG. And there are others around this table that have a lot of very intimate experience with this sort of standard of where we are

currently. So I encourage you all to sort of speak about this in terms of your experience about trying to implement these measures across various different communities.

So the challenge is that when you interchange these measures with various different risk adjustment models, so as an example, if a community was trying to use this measure using DxCGs, for instance, they would not be able to call these NQF-endorsed for the reasons that the measure with its various different risk adjuster hasn't been tested for reliability and validity.

And so this introduces a significant amount of burden or barriers for various organizations across the rest of the country, since they have already introduced and used -- invested in these various different commercially available risk adjustment models.

And so this was something that was discussed during the first cost and resource

use project and in some ways is actually not very unique to cost and resource use, as Karen will describe later on, and the fact that this is a very similar issue to the fact that when we look at on the quality side, there is also an investment in terms of risk adjusters.

So if you can go to the next slide. There is a lot of rationale in terms of allowing a single measure with multiple different risk adjusters that have been presented to NQF in terms of a need primarily that allows markets and users to be much more flexible in the risk adjustment model that they have already purchased. Single markets generally have commonality around a specific risk adjustment tool and transitioning to different risk adjustment tools is really inefficient in terms of opportunity costs, licensing fees and the like.

So the challenge here is that if we are looking for national comparisons of being able to compare Minneapolis, as Jack

described, Minneapolis in Peoria, that might be the two cities of this group for today, that you may potentially need a single tool for those comparisons and those that are actually using these tools would understand that you can't use a measure with two different risk adjusters and try to compare the results. And that would be broadly known by those that are using this measure.

And the challenge of comingling the results would actually not really be that much of an issue because of those that are using these measures would understand that the measure results are not necessarily comparable. And we don't really need to revisit or create new measures that have the same specifications, just have different risk adjustment models because that would introduce a whole series of new measures into the portfolio with only a single difference.

And that has been the rationale that has been provided to support multiple

different risk adjusters. And I encourage, as we get into -- we will have time to discuss this, to have comments on this particular issue and I encourage all those who feel strongly about this issue to discuss it.

And so before we get to that, I want to Just also turn to my colleague Karen Pace to describe some of the NQF principles and the relevant criteria to these questions around allowing multiple different risk adjusters in a single measure and this issue about comparability.

And again, I would encourage all of us to consider the fact that this criteria that we use to evaluate cost and resource use measures are the same criteria that are used to evaluate quality measures, which is why we want to make sure that we are consistent and this issue actually has much more ramifications then just in the field of cost and resource use measures.

So Dr. Pace.

DR. PACE: Hello, everyone. So I know you have had lots of discussion over the last couple of days so thanks for hanging in there with us.

I am just going to mention a couple of things to keep in mind and as Taroon said, this is not just an issue specific to resource use measures. That risk models are used with outcome measures and we actually have encountered the issue of trying to measure the same thing with different risk models. And to the extent possible, we want to be consistent across all of our measure evaluation.

But just to lay the groundwork, as you know, NQF endorses national standards for performance measures that are intended for both accountability and performance improvement. In order to be useful to make conclusions about performance, especially relative performance, all entities need to be measured exactly the same way. So there is a

couple of things here. One is standardization and one is national. And so we need to think about that, if there really are unique differences in some of these things regarding resource use measures.

Typically, we seek to endorse the best from among competing measures whenever possible because this minimizes the confusion created when the accountable entities are scored and ranked differently when performance measure specifications are different. So the two measures you were talking about this morning, if providers get different scores and end up in different rankings, based on those two different measures, we have introduced a lot of confusion in the measurement space. So that is one of our main concerns of why we try. And as you know, it is difficult and not always possible to do that. But that is our goal.

And to the extent possible, as I already mentioned, we want our criteria to be

broadly applicable across performance measures. And obviously, there are going to be some things that are unique to certain types of measures. So risk adjustment is only a concern for outcome measures and resource use measures. But to the extent that we could have consistency about risk adjustment for those types of measures, it would be a good thing.

And so in terms of the criteria that are relevant regarding the issues that we are discussing is first of all that we do think they should have a risk adjustment strategy and that is specified. And if multiple data sources or methods, and this is where risk adjustment comes in, are specified, there should be a demonstration that they produce comparable results. So this gets to the single measure issue. If we have a single measure that says you can risk adjust this way or this way or you have five choices, first of all, we don't necessarily have a standard

measure. And second of all, if you applied those different risk models to a group, would you get comparable results?

The next one is validity testing. So again, one of the propositions posed earlier was that the only thing you would need to review is the risk model metrics across these measures. But I think it is still an open question is would you really get the same validity results if you applied different risk models. Maybe so, maybe not. The same way with reliability. Would you get the same reliability results?

So I am not so sure it is just as simple as just looking at each of the risk models in isolation when we are talking about performance measures. Next slide.

Okay so I just wanted to kind of present the backdrop in terms of NQF principles and criteria in terms of why we are grappling with this and wanting to have some discussions with you about it. And I think I

will let Taroon introduce this next discussion because, as he said, this has been introduced into these conversations.

MR. AMIN: So I will turn it over to Syed in a moment. And again, I appreciate you traveling here and being here with the committee today. And the fact that there was a 2007 analysis of health risk assessments that really looked at the differences between the various different risk adjustment models, which has been cited a number of times by the first committee that we were looking at, looking at the Society of Actuaries report and has been cited again also by many members of the community that show -- that argue that this report demonstrates comparability across various different risk adjustment models, which is slightly different than the question that Karen has raised, which is around the comparability of the measure score, which is what we are really trying to understand here.

So Syed if you would please just

give us kind of a sense of the work and you assessment of this question as it relates to the work that you have done in the past.

MR. MEHMUD: Sure, thank you, Taroon. And thanks for inviting me here today. I think I had a chance to hear some of the conversations around the table and I will keep my comments brief because I know I am talking to a very informed audience here.

I think we did this work a while back, this specific report I don't know six years ago, gosh. And so we looked at several different adjusters that were available in the marketplace at the time, I think about 12. And basically ran a lot of tests on those adjusters that were mostly around accuracy, statistical performance and that sort of thing. And the tests, one of the metrics that received I guess more air play than the others was the R squared metric. It is one aspect, I would say of a risk adjuster and I have dealt with risk adjustment in Medicare and

Medicaid, commercial, all settings for a long time and I know that that is only one sliver of the different dimensions of a risk assessment tool and a risk adjustment tool and that is one aspect of it.

The report also talked about predictive ratios, which is a pretty important concept, as well. And that is basically what is the risk assessment tool predicting, for example, the score to be? And the score is, in fact, just an estimate of cost. That is how these models are developed. And divided by the actual.

And so that gets you a little bit of a sense of whether there is any bias in risk assessment predictions. And there is a bit of a truism -- not a truism but a saying amongst risk assessment practitioners that sometimes it is more important to not be biased than it is to be accurate. So I think the report, this particular report talks a lot about accuracy but bias is also an important

consideration.

And in this report, as I mentioned there were several different models that were tested. The models -- I had to actually print it out because it has been a while so I had to read, just glance at it again. The methodologies employed by the different tools, I was going over it this morning and I was realizing that they were more diverse than I had thought. Because over time, I think since then it is kind of you tend more towards a certain kind of developmental methodology like how you build these models and regression is something that comes up a lot. But then these models, they employ, to some degree, different methodologies. So that is not the only technique in town.

So I think that is kind of a little bit of a context in the report. So coming back to I think what you are alluding to, can you use different tools, these different risk assessment models, to develop

something and then compare that thing. And I have to preface that by confessing that I don't know enough about these specific measures that you are discussing today but I do know that you are using risk scores to adjust those measures in some way in order to make them comparable.

And so I guess a real question then shifts to are the risk scores comparable, either on an individual level or in some kind of group level, maybe by a provider or health plan, and so forth.

And I think that there can be different sources which might cause the scores to be different amongst different tools. So if you take different tools and apply them to the same population, you may not get exactly the same score from it. And one source is what information is used to develop those models.

So if one model is built using a particular commercial large dataset, another

is built using a different dataset, another is built using a different dataset. The meaning of a 1.0, the meaning of a score becomes different. So the score itself is not directly comparable at that very kind of like basic level.

The other differences might be due to how the groupers developed, how the information from ICD codes or NDC codes or any procedure codes or any other information, how that is used to develop clinical markers and what rates are assigned to those markers.

Now the development of the clinical markers is a clinical aspect to that and if those are different, they can cause differences amongst risk assessment tools as well.

So for example if you have one assessment tool might break up chronic conditions such as diabetes and to fight different severity levels, another might do just one or two. And that would make a

difference of when you apply different tools to the same population.

Then the next source becomes how do you develop weights to assign to these different clinical markers. And as I mentioned, regression is one approach but there are other approaches that these tools have used. For example, there is the actuarial cell approach where you have mutually exclusive condition categories that you develop weights for. There is the episodic kind of an approach where you develop episodes and see how care was being delivered to a patient and then assign based on those episodes. There is a hierarchical approach where you develop a certain set of markers and then you collapse them down, based on certain hierarchical rules-based kind of logics and model-building process.

So I think through all of these different sources that might cause you to have different risk scores if you take two

different tools and apply them to the same population, there has to be some, obviously, testing and validation to understand the magnitude of these differences.

And I think I heard a point earlier that on the individual level they can be different but maybe on a group level they aren't so different. So I think that is also a good area for validation and testing.

The one thing, one short comment I would make there, is that any risk adjuster, you could take the same risk adjuster and you apply it on an individual level and you group those risk scores at a group level, the differences amongst risk adjusters become smaller and smaller.

So this particular report, for example, that we did in 2006 at the Society of Actuaries was measured performance on an individual level. So when you take these tools and you apply them on individuals and you measure accuracy on an individual level,

these tools are, they produce differences amongst accuracy metrics and so forth.

If you had done this on a group level, there would still be differences but the tools would get a little bit closer together. But is that close enough? And I think even though the risk scores themselves might be closer together, they are getting applied to a larger population. And so I tend to come from a world where we do -- so we are talking mostly about risk assessment here. And I come from a world where we are very heavy on risk adjustment, which is when you are actually changing money, or money is exchanging hands based on the calculated risk score, for example, in a captive market and so forth. For example, as in the ACA in 2014 or in the Massachusetts market. So there, when money is exchanging hands, it becomes really important, even small differences in group level risk scores can be significant, can be material and can make a big impact to

organizations' bottom line in a risk adjustment environment.

So I think it needs very careful testing to understand how comparable these are. But you know my opinion is that there will be differences and differences could be attributed to the data used to build the model, the modeling, the model-building methodology itself. The weight assigned to methodology to the different clinical conditions.

MR. AMIN: Thanks Syed. Before we get into the committee discussion, obviously there is a lot of opinions on this topic. And so we have various different experts, again on the phone. Many of them come from measure development worlds. So if it is okay with the committee first, maybe we can go through a 15-20 minute sort of a comment period but a little more structured, and kind of get some assessments from those in the room and those on the phone.

First the structure and the basis of this conversation, so making sure that that is agreed up that where we are trying to go in the needs in the community in making sure that that is clear. And then some general sense of kind of where we are trying to go.

Do you guys have questions on that or comments? Okay. All right, well it is obviously a lot to talk about here.

So I mean I can go through at a high level what the discussion questions are. But actually I would like to wait to see if there is any sort of framing comments that others on the phone or others in the room have before we get into sort of specifics. Go ahead, Sue.

MS. KNUDSON: Thanks, Taroon. So Taroon did a nice job of outline sort of the issue we had and Syed's comments are very helpful, too. And I just wanted to comment briefly to kind of take it up a few thousand feet from a practical developer end user point

of view.

So you know someone earlier in today's discussion mentioned that we have all these commercial ways of doing this. And I sort of characterize that to a little bit of unneeded chaos in the measurement system. And so as we, as a developer, who came through this process a couple of years ago, as we sort of observed others who had similar measures like the IHA in California, the only difference is the risk adjuster, are working now with uptake across the country, many people ask us well can I implement this because I am using ERGs in this community or DxCGs or what have you. And the answer has been just as Taroon had mentioned that you can only call it NQF-endorsed with the AGC tool. So we were thinking really more from a practical point of view, we could come through this round and only retest our measure using a few different risk adjusters. And really the spirit in which we did that, because

frankly we all have day jobs, too, but really we are committed as an organization to more affordability, sustainability in our country. So rather than have all sorts of new measures come through and then we have these harmonizing discussions, we were thinking a more rational way to do it would be to test the different risk adjusters and then you, basically have the same guts of the rest of the measure being all the same. And so that is the only difference.

And that, to me, is more of both end strategy. We are, at one time, working to meet local markets where they are at and at the local market level, this is tied to money. And so when we have markets invested in the use of a different tool than what we have endorsed, like California, for example, when they have spent all this time and energy getting accustomed and comfortable across all the stakeholders with a tool, you don't want to increase all those burdens that Taroon

outlined and sort of upset the whole apple cart. At the same time, we have the need for a national study from a policy point of view, from a payment disparity and resource use disparity and understanding those things from a national view. So that is the other side of that both end strategy. And that is, for those of us who do this work in the field, we know, we basically have to use one risk adjuster to do those national comparisons. That goes to the work that we are trying to get off the ground with the Dartmouth on a commercial companion to the Dartmouth Atlas using this measure.

So that was more of the practical point of view. But the NQF process, particularly that section which was outlined 2b something or another, where it calls for the comparability is I think really where practically we bumped into, even though we could test these, and we have tested at least two other risk adjusters, we have got a write-

up done on one of them. You know they do produce reliable and valid results when we just retested them, the measure at least in those two.

And so it was really in that spirit that we wanted to meet communities where they are at and at the same time know the difference when you want to study national comparisons and make policy decisions and other more macro decisions and inform health policy based on that.

So those are just a couple of context questions or comments I would make as well.

MR. AMIN: Are there any other comments that are on the phone?

OPERATOR: To make a comment at this time, please press *1 on your telephone keypad. There are no comment or questions.

DR. NEEDLEMAN: I think the order of the things -- you asked if we had any comments about the scope of the discussion.

I think the order on the chart there is not quite right. I think the second bullet point about what would be a demonstration of comparability is actually the place to start this. I am blanking on the name of the person from HealthPartners who just spoke. But I think the word there is we have got reliability, variance estimates, validity. The real issue here is the last word she used, which is consistency. When you apply the same --

Yesterday we were looking at measure and you split the sample in half. The data, there was enough variation in the data that 30 percent of the folks that were in the top quintile or the bottom quintile dropped into a different place. That is a data variance. What we need to understand is whether if you take the same data with the same ultimate aggregation but a different risk adjuster, how consistent are the results from one across the different risk adjusters if

you get the same ranking of people, if you get the same location of people, how similar or different do those need to be? If the results are consistent, if you come to the same conclusion, and I think one of our colleagues here talked about we use different measuring and qualitatively they tell us the same story and we move forward with it. That is the is the issue.

And I didn't hear Karen and I didn't hear Syed talk about whether anybody has done enough study with the same data on patients to ask about the final end measures, as opposed to looking at the R squares or the equivalent measures about the individual measures. That is the comparison we need to figure out where to go. And I heard that alluded to by Sue. And Karen you look eager to speak.

DR. PACE: Well I just want to say, that is the whole basis of the criterion that we have that if you are going to specify

a measure and have alternative methodologies that you demonstrate comparability, which is for consistency --

DR. NEEDLEMAN: So I see this discussion is basically saying can we find from some standard measures some standard stuff enough evidence that whether you use ACGs or DxCGs or one of the other things that all produce about 20 to 23 percent R squares, whether there is enough comparability we don't need to test it in each new measure. We can say if we have approved it with ACGs where you can basically run it with these three other measures and we are comfortable you will get the same results.

DR. PACE: Right but I think that we wanted to also establish some common understanding because people have been alluding to, if they have the same R squared of the risk model that you will get the same performance scores.

DR. NEEDLEMAN: No, I don't

believe that.

DR. PACE: And we want to establish some common understanding that that is not the case and that we actually do have to do these empirical analyses.

DR. NEEDLEMAN: I agree. And we have been having this conversation in the field for a decade now and I am a little shocked that those studies have -- we don't have piles of those studies that already can tell us the answers to that.

MR. AMIN: So actually before we get started on the conversation, maybe it is helpful, we set up a few questions that are related to -- so if we can go to the Society of Actuaries discussion question. And this is going to the exact point that we are having here.

The first thing, and again, we are not, this is not just a HealthPartners issue. We just want to make this clear. This was raised by a number of different groups that

are interested in using -- and the same issue around the multiple methods, the fact that we allow these options within resource use measures, which was a strategic design of how we move forward with this first effort. And what we are seeing now is some reaction to that. And that is why we need, we as a committee, need to be accountable for those decisions and kind of where we were in terms of our assumptions. Meaning that we wanted to allow the flexibility in terms of various different components of the measure.

But specifically, let's go to this issue, which is that Syed did a nice job of describing kind of what the Society of Actuaries report described. And again, what we are trying to do is make sure that we are clear not only within this room but also making sure that as we write this up and provide some guidance back to the field in terms of what we are expecting, that there is this issue, which is that when you look at

various risk adjusters and they have similar performance metrics in the sense of R squared, what does that say or what does that not say about whether the performance scores for accountable entities are actually comparable?

And whether overall reliability of the performance score with different risk adjustment models will be similar, what does that tell us? Again, Jack, this may seem to some people that this is -- well you know, we have been doing this for a decade. It is not generally agreed upon in the field that using these various different risk adjustment models with "similar performance characteristics," which is --

DR. NEEDLEMAN: Yes. My point was not that it is obvious from the R squares. it is not obvious from anything in the Actuaries report, as well done as it was. And it is extremely well done.

The issue is the next step has not been taken and it could have been taken five

years ago, ten years ago, take a validation dataset, the same patients, the same distribution of diseases, the same distribution of use, the same distribution of cost, the same measure and run it with different risk adjusters and see how similar or different they are. That is what we really need to answer this question. It is disturbing to me that we haven't done that analysis yet and we are still talking about mights and maybes, and we might expect and we should expect. That needs to be done and it should have already been done but if it hasn't been, it needs to be done. I am a crude empiricist here. I want to know what difference it makes in the rankings and what difference it makes in the estimates of the values.

MR. MEHMUD: Yes, if I can make a quick follow-up comment on that discussion. I think partly the reason it hasn't been done is because generally in any risk-adjusted

market, you don't have different risk adjusters being used. You have the same risk adjuster being used. And so what really, you know at six years ago, whenever, when risk adjustment was kind of like really coming into the fore, folks were more concerned about what risk adjuster should I go with for this application or that market, or that program. And so they were wanting information on things like statistical metrics, inaccuracy metrics and so forth.

But if you are trying to have a more general application with the understanding that different risk adjuster models may be used, then that question becomes extremely important. And I would say that that is a very specific application of risk adjustment.

And that, to my knowledge as well, that study has not been done where you kind of look at different tools and measure their output. And to me it is not so much -- I

guess I am little vague on what performance means but the more basic question is you take the same tool, you apply it -- sorry. You take the same dataset and you apply it across different tools. Do they all produce the same result, the same output on an average level for an individual that won't perhaps, but if you aggregate it up, do they produce the same answer?

And in that study, I am taxing my memory a little bit, but in 2007 we used 12 different models on the same dataset. And in order to compare them, we could not compare the output as is because you average the way that you run these comparisons is that you have to calculate the R squared metric and so forth. But if the risk models are producing scores that average to different levels, then your comparisons will not be valid.

And so for example if one model, you run say a million people through a model and one would hope that the average score is

1.0 but it never is. It could be 1.05, for example for that model, a different model you run it could be 1.07. A different model could be 0.93. So we have to normalize for those differences across all those different models first and then calculate the accuracy metric so that at least we are taking that element and that was the first source of variation that I described is the data that is being used to build the models up. So that is what will need to be tested.

And that would be a very worthwhile study if the application is, once again, in that very specific application where you are considering the use of multiple risk adjusters in order to compare some metrics across different populations and datasets, and so forth.

MR. AMIN: Bret?

DR. ASPLIN: Maybe I am thinking of this too simply because I am certainly not a methodologist who can dive into the risk

adjustment like others in the room. But I do think this is, practically speaking, it is just going to boil down to the harmonization process and groups like us reviewing both qualitatively -- Jack's question is is it telling the same story, this measure with multiple adjusters or not -- and then quantitatively measure by measuring harmonization. And then over time, developing guidelines for just like you have other guidelines in your harmonization process now, developing guidelines for risk adjustment in the harmonization process. I don't know how you get around groups like this, measure by measure asking the same harmonization questions we already do with different risk adjustment models.

So I think one of the challenges is not only some of the questions that have been raised about how you do this study but who are you going to put the burden on to do that? Because it can't go on the original --

the burden to be a measure developer has already gone up by such orders of magnitude over the last five years that -- and Sue just I think pointed that out very nicely. I mean you can't expect HealthPartners to come back and do all the analysis and testing. Maybe they have done some of it, but to get through the initial bar to a committee, to present the measure with risk adjuster A and then oh, by the way, here are three different risk adjuster models that we have tested. It is not going to happen.

And so if these other communities that are using different risk adjusters want to stick with their adjuster and have an NQF-approved measure I think the burden is on them to do some of the testing and bring it forward for harmonization discussion. Now will people do that? I don't know.

MR. KITCHING: If I could interject, Taroon?

MR. AMIN: You can. Just get

through maybe these three and then, if you don't mind. Thanks.

DR. GIFFORD: So of course I am not a methodologist but I am a doctor so I know what I am talking about on the statistics of R square.

(Laughter.)

DR. GIFFORD: I am just going to tell you I would agree very much with your comments and expand on them a bit in that there is no risk adjustment that is perfect. Many times they are not even close to perfect. And I am struck by how many providers and groups out there, as long as you say it is risk adjusted, they suddenly seem satisfied with the measure, even if there is just one variable in the risk adjustment model.

I think your suggestion makes a lot of sense on the burden. Is that, we are proving it with a risk adjuster. If others want to use different risk adjustment, that is fine. There is no pragmatic reason they want

to use it. It probably doesn't really vary that much in the grand scheme of things, but if you are going to compare providers across different areas with different risk adjustments, you can't do that. I think you need to pick one and use it. And then if it is going to be picky and say it is NQF-endorsed, it has got to come through and get reviewed and evaluated in a rigorous manner. If others want to come through and do it, that is fine. If they want to use it and say we are basing our measure on NQF but we use a different risk adjustment because of X, Y, and Z and they can justify it to their community and their community buys it, have at it. I wouldn't discourage it and I wouldn't say it. But to have multiple -- for us to be endorsing multiple different risk adjustment, that just doesn't make sense to us.

MR. AMIN: Nancy?

DR. GARRETT: So as someone who has used some of these different models, I

think one of the things that is important for everyone to understand is that these are commercially available models with a lot of complexity that do not -- they are not open source. And you can have a couple of different analysts, even in the same organization implement these groupers and get different results because there is so many different decisions about how the data is prepared, what the settings are that you are using. And to have this goal of consistency, I don't know that it works in this case with the way these methods work.

The way to get consistency would be to have one organization calculating the data and running it through the same risk adjustment grouper and using the exact same methodology. To try and get consistency I mean the manual would be 100 pages long. And by the time it was written, the ERGs would be onto the next version and things would have changed. So I don't know that it is really

attainable in this case.

So I think the important thing is that there is risk adjustment. But like you said, David, you have to be careful about the comparability and when you are comparing across organizations using different methods.

And a lot of times, these measures are going to be used for comparison over time within the same provider. And so what is really important is that the methods are consistent over that time period to be able to understand how things are changing.

MR. AMIN: Chad.

MR. HEIM: Thank you, Taroon. So just to kind of follow-up with Jack had said and Brent had said and Nancy had commented on. Consistency is the important thing. So we had to do some testing on additional risk adjusters since we actually have them in-house. And actually we are kind of preparing for possible submission to NQF for this very committee to take the technical spec as

written. To get at Nancy's comment for the consistency where the truncation levels are set, nine months' requirement for enough diagnosis codes and just to follow the technical spec that this committee approved and just swap in the different risk adjuster and see if we see any major movements in terms of at the provider group level. And we have found some pretty consistent results when we are actually swapping in different groups.

So I think the key points are the consistency of the technical spec that NQF improved endorsement that kind of aligns going across risk adjusters, because those are critical decision points. But then we found consistency. There was a lot of providers that were if you were high, you were consistently high, no matter what group we were dropping in.

DR. WEINTRAUB: So I think I will tell you a story from what goes on in real life in hospitals. Because I think that what

is going on in the commercial space with risk adjustment is a mess.

We found that when are looking at our heart failure patient population that we had using a commercial risk adjustment, and we seemed to be getting very bad scores. And I didn't understand it.

So I went and looked under the hood at just what was going on and the vice president of our hospital in charge of this said, it is risk adjusted. There is no problem here. But I went and looked and it was a joke. And it was very easily understandable to clinicians why it was a joke and why our risks of mortality in heart failure seemed to about double with risk adjustment. We would have been much better off just looking at it unadjusted because in fact it was doing something that was absolutely wrong.

So I think very good risk adjustment is critical. And I think this is

a great place and vital place for NQF to be.

Now how to compare risk adjustment. I think that an R square along is not the best metric or for discrete variables, a CNX, there is much more we should do. Clearly, calibration of risk adjustment is very important and external validation, preferably with another data set.

And while it is hard, there is a lot of work going on comparing risk-adjustment methodology, I think this is a very worthwhile space for NQF to be in. If it is not going to be NQF, who is it going to be?

Now finally, I think that risk adjustment applied to the individual patient is not particularly been the space of NQF, at least it seems to me. But it is very important how we function every day because more and more we are developing and applying prediction models at the level of the patient. And very often at that level, they are not very well validated.

MR. AMIN: Larry?

MR. BECKER: So as a consumer, there is another thing that is really important and I think that is transparency. And when I hear people talking about it is not open source, I cringe because I mean, so I'm sure consumers can't understand this stuff but experts can. And so there needs to be some endorsement of this was done according to a transparent methodology. You can agree with it. You can argue with it but at least you can attempt to understand it and you have some sort of confidence that somebody doesn't have some twist on it for their own purposes. So transparency is really important.

MR. AMIN: Gene?

DR. PENSON: I have a -- oh, sorry.

MR. AMIN: Just Gene first. Thanks.

DR. NELSON: A really good conversation so far and I think going down the

table here what Jack is recommending that you actually have data to be able to do the calibration and what we here at HealthPartners are trying to do is exactly what would be called for to get the PROsetta Stone that allows you to go from one risk adjustment system to another and to do that in a way that is accurate and calibrated.

So the second point would be that I sit in the hallway, literally in the middle, and on one end is Jack Wennberg and the other is Elliot Fisher. So I am a little bit like Forrest Gump in the middle.

So I have papers from the New England Journal of Medicine 2010, JAMA 2011, BMJ 2013, and there will be another one coming out pretty soon. And they all show that there is a fundamental flaw that some of you probably recognize in all of the risk adjustment systems that use claims data or administrative data that because it is confounded by the intensity of services.

So literally, one of these papers shows, JAMA 2011, that if people go from a high-intensity area, to a low intensity area, lots of doctor visits, lots of hospital stays, lower hospital visits, lower hospital stays per capita, if you go from high to low, you start to look better. Your risk adjusted scores are better. And the converse is true as well. If you go from a low to a high area, Minneapolis to Miami, you are looking sicker faster. And so there is a fundamental problem for both mortality prediction, cost prediction, resource use prediction, et cetera, across perhaps all of these systems. And so that puts us in a bad position if we are trying to look out for NQF's interests in national standards for performance and accountability. And so I would just start to think that this brief conversation that we are having now really needs to be done in a very rigorous way and looked across all the measures.

Bob Brooks wrote a little commentary a while back and it was labeled, in effect, quality is dead, long live value. And as you start moving towards needing to have the building blocks to measure value, you had better get the building blocks right. And risk adjustment goes across a lot of these building blocks, outcomes, and costs. Not so much technical quality. Good for that.

But there is a very big issue here. And there is a lot of expertise out there. And the evidence is building up that this risk adjustment building block based on claims data has some inherent flaws. And the good news is, there is probably solutions that they could be improved, using modifications.

MR. AMIN: Okay. So before we continue on in the discussion, I am going to ask Evan to just go back to the criteria slide and just put up a straw person for just to react to, which is the fact that basically we have heard a lot of need from the community

around various different pieces of this. And the straw person basically is to continue with where we are right now, which is that if you have specifications that include multiple methods that you have to demonstrate that they are comparable and that every measure, obviously, would still go through our criteria, which is to require testing and the various different components of our risk adjustment strategy criteria.

So I guess that is where we are landing. That is what I am hearing. That was pretty clear from Brent and Jack and others. I mean, it couldn't be any more clear, at least from them. I mean that in the best way possible, obviously.

So if there is others that feel or want to say other things related to that, I think it is important to say your opinion here. You know, I am also really interested to hear from those that are in the community, again, to make sure that your perspective is

heard here. And so we will start with you, Don.

MR. WOLFSON: I was just wondering a simple minded solution is to get a set of approved risk adjusters that NQF approves. And if you don't use those, then you had better come up with something that would justify that. Why can't there be whatever number of a pre-approved risk-adjustment methodology?

MR. AMIN: Karen, there is a few, I think, areas that we could take that. I will start with a few and then, Karen, you jump in and correct me because I am sure I will go off somewhere.

So there is a few, which is that actually it kind of goes to the foundation of this conversation which is that we endorse performance measures. And just because even if we did say that various different risk-adjustment methodologies were appropriate and "endorsed," that may not actually tell us

anything about the output of the performance measure. So that if you are actually looking at they are great risk adjustment models but the specifications are inappropriate, and so at the end of the day what is that really telling us? And so that is actually some of the challenge here, which is we are endorsing the end piece, which is what we care about, which is the end score. So that is the first issue.

The second is the fact that there is proprietary interests here. And that is reasonable. I mean obviously we accept components, proprietary components within measures. But we don't want to -- we probably as a collective community may not want to get in the business of endorsing one methodology or multiple different methodologies because at the end of the day you still want to have this discussion around harmonization and picking a best in class.

And so in the scenario where you

do have various different measures that include different proprietary components in them, meaning competing risk-adjustment models, if you will, shouldn't we -- again, this is sort of the foundation of what we do here is to have a national standard. And if that national standard has only one proprietary -- is differentially benefiting one proprietary interesting, I think we are going to find ourselves in a difficult position.

But Karen if you have anything on that specifically.

DR. PACE: I think mainly that currently NQF endorses the performance measure and all of the components that go with that. And it is something that certainly if that is a recommendation, it can be explored in terms of NQF endorsement. It is similar to does NQF endorse, for example, instruments and scales? We have had this discussion in our recent patient-reported outcomes work. NQF doesn't

endorse the variety of depression assessment scales but we are interested in endorsing a performance measure that talks about percentage of patients that actually have a remission of depression. And obviously, the two are closely related in that case as well.

So it has broader implications. You know, it is certainly something we have heard before. And Helen, I don't know if you want to comment on that in terms of our current endorsement process.

DR. BURSTIN: I don't really have much to add. I agree with you and Taroon. This is just really complicated.

I mean, Daniel, I completely would love to just make this work for as many people as possible but then you get to the concerns that Bill raised and others about just fairness. And I think the measures aren't comparable putting them out there in that way. It is just a question of whether we are really fulfilling our responsibility.

MR. AMIN: So Syed, Dolores, Lisa, and then we will open it up to comments in the room and on the phone and then lunch.

So, Syed.

MR. MEHMUD: Well I will not delay lunch. I will be very brief.

I think just hearing this conversation, obviously the emphasis on transparency and making the measure as applicable in as many situations as possible.

One quick mention that I would make is that 2014 will probably see one of the largest scale applications of risk-adjustment ever with the ACO risk assessment model being applied in all the states, except one, that I am aware of in the small group and the individual markets. That model is going to be ultimately available. And right now they have published details of it. They haven't made the software available yet but perhaps that will be. So that is one thing to keep in mind is that next year there will be that model.

Everyone, pretty much everyone is going to be using it.

MR. AMIN: Syed, and that is based on the HCC approach calibrated in the commercial sector?

MR. MEHMUD: Yes, I think they took the approach that they follow for the CMS-HCC -- oh, well, he may have left. But they took that approach and they applied it to a commercial population. It is a different model than the CMS-HCC but it is the same sort of philosophy of model building that was used.

MR. AMIN: Okay, Dolores.

MS. YANAGIHARA: So on the comparability issue, I think there is different ways to talk about comparability. I just want to make sure that all understand what we are talking about there. And that maybe all three different ways that I am conceptualizing it.

So there is comparably valid and reliable for different methods. So is each

method kind of comparably valid and reliable. And then there is, if you are applying different methods to the same providers, do you get kind of a similar rank order, so to speak? So that is another way to look at comparability. And then there is comparability if we are using different risk adjusters on different populations is it comparable?

So there is three different levels of comparability and I am not sure which ones we are talking about on criteria 2b6. Is it all three of those that we are talking about?

DR. PACE: In 2b6 we are really talking about that they produce comparable results, meaning the score, the performance score that is given to the accountable entity, whoever is being measured.

And regarding you first that each method or each method results in a reliable and valid performance measure, I just want mention that our discussion here does and will

-- and that is kind of the next set of questions, doesn't necessarily preclude that we could endorse multiple measures with different risk models, with appropriate justification. I mean I think that gets into a lot of other issues but it is something that we will be asking you to reflect on.

So I think the first thing when you mention each method can produce a reliable and valid performance measure, even though there may not be comparable results, that is the question of competing measures. We would have to see them as competing measures and look at each individually. Do they have the same kind of reliability and validity.

And that is another question about how far down that road we want to go. But I think that is another question.

MS. YANAGIHARA: Yes, I think that the difference that I am trying to get at with the second -- and thank you for those comments -- for the second and third things I talked

about is you can have comparable results within a data -- like the comparable performance, relatively speaking. But when you start comparing on an absolute score, it might be different.

So in order to compare across different populations using different risk adjusters, the score has to actually be comparable too. Right? And so are we talking about both of those?

DR. PACE: We are interested that if the measure is specified to measure the cost at the group level for all their patients, that that risk model, that the two different models -- so all of those other specifications are the same in terms of who is the accountable entity, the level of analysis, the patients included. If you apply two different risk models, would those providers be ranked comparably?

If you start talking about different populations, then we have other

questions about again, whether you can have one measure that adequately addresses different populations.

MS. YANAGIHARA: Yes, because I think that that makes a big difference. Because what I heard Chad saying is that if you take the same population and you apply these different risk adjustment methods, they track pretty well. So you get pretty comparable results. But then I don't know that that necessarily means if California is using DxCG and Minnesota is using ACG is that we are going to be able to compare our results. So I think they are two different things and that makes big implications of what we are talking about here.

And I almost think if this is a national arena, that we need the third thing is really what we need, which is really -- and then you have to have one risk adjuster. If you really want to be able to compare across the country, you need one risk adjuster. And

to me, that sort of cries out for a publicly available risk adjuster that performs sufficiently well, like equally to the proprietary ones or better. And so I don't know how the HCC one that is being developed for the exchanges would work. But might this be something that this committee could call for is a publicly available risk adjuster that performs well. I mean, that would then be available for everyone to use. And that might be a recommendation that could come out of this committee that would be very helpful to take to other different organizations and entities to really grapple with.

MR. AMIN: Lisa.

DR. LATTS: Well I think sort of the elephant in the room is that these are proprietary systems and that there is a revenue model and a business model associated here. So there are several organizations that their model is to show that their risk adjustment methodology is better than the

others, so they can sell it. So there is no interest in there being a public model. It is the proprietary model and showing that theirs is better.

So I think as long as that exists, we are not going to get to what this committee is asking for because there are entities that are looking to make a profit off of this. So I think that needs to be part of the discussion.

MR. AMIN: Okay. So, are there any comments on the phone? Operator, are there any comments on the phone?

OPERATOR: To make a comment at this time, please press *1. There are no comments or questions.

MR. AMIN: Are there any comments in the room?

MR. BANKOWITZ: Yes, I do have a comment. Thank you.

I am Richard Bankowitz with premiere. Let me say in the interest of full

disclosure, premiere does have a risk-adjustment methodology that it works on and uses in its products.

The first comment is about trying to compare these things. It is really not as cut and dry as providing saying one is better than another. I have done these types of studies and if all the time you can line up risk adjuster A with risk adjuster B, there usually is a diagonal line at 45 degrees but some will over-predict and some will over-predict. And it is a question of tradeoffs. So one might in a different population. One might be better in population B. It is not quite easy to say A is better than B. That is the first point.

I think what really would be helpful for developers would be if NQF would lay out some best practices or guidelines about how to handle some thorny issues. So two issues, one I hear about every time I am at a meeting on risk adjustment and one I

basically never hear about.

The one I always hear about is how are we going to handle socioeconomic status and all of the things that go with it. So obviously, if we have patients who have poor access to care, have poor healthcare literacy and have poor access to nutrition, that is going to make a difference.

Do we model that in or not? We can debate that. We had a discussion yesterday it seemed clear in the population of dual eligibles. We don't want to model in different approaches to care.

So if you believe they cost more because providers are just over-utilizing, then don't model it in. If you believe they cost more because they actually are in some ways sicker, then you model it in. So I think NQF could be more explicit about how to handle that.

The thing that I actually never hear discussed is how do you handle secondary

diagnoses that are caused by the healthcare system, complications and the like? So I can build a model incorporating pulmonary embolism in elective surgery and that model will always have a better R squared and it will always have a better C statistic than one where I don't model it in. And I would argue we don't want to model that in because if we have a patient who is off the chart because of pulmonary emboli and length of stay cost for mortality, we want to find that patient and not risk adjust it away. But there is really no guidance on a best practice because some risk adjusters include everything and some make an effort to model those things out. So that will be very useful.

There are other things, too, like just the patient population. Mortality has been going down year after year. If you model a population from 2008 you are going to get very different results from modeling a population from 2012. So the breadth and the

dates of the population could be helpful.

The way we handle chronic diseases, that is usually not terribly important, although that is what we spend most of our time debating. But things like outliers and transfers, how do you handle those? What is the best practice.

And then lastly, end of life care. And depending on whether you are modeling inpatient mortality or you are looking at cost of a population, you have to figure out how you want to model that or not model it because it makes a huge difference. So if NQF would have maybe some guidelines or convene an expert panel or summit or to help set best practices, that would be very, very useful.

MS. KNUDSON: So I just had one last comment and really ask for a clarification and a reconsideration of the process because I think based on Syed's comment, we would support, too, that the best case scenario is to have an open source risk

adjuster that accommodates the needs of a commercial population. But absent that, what we have is the best we have and those are commercially available risk adjusters.

HealthPartners has said we would test our already endorsed measure using that publicly available one when it becomes available. What is difficult, and Taroon you teed up a measure steward burden question that we really didn't get to or maybe just around the edges, we would test that on our already endorsed measures. So from a practical point of view, we would ask for NQF to reconsider the process that we wouldn't have to then fill out the application, do the full vetting of all of the rest of the spec that is already endorsed, use another Steering Committee to go through the entire specification again which is already endorsed, but rather only look at the specific differences in the reliability and validity testing results.

And I know those sections include

more broad components that also apply to the spec but I would say we would just be very precise with what is different about the results of this open source. Because absent that flexibility in the review process, we won't have the resources to do this on behalf of the healthcare community in this country. But we would, because we are very invested in affordability for the country, we would do that testing on our already endorsed measure. We would just ask for flexibility in that review process to be very precise around what is different with that testing and not require us to re-vet the entire measure again.

And I don't think -- I think based on the discussion, I just wanted to make that ask because I think it is very practical based on the discussion. But I just wanted to be clear about the ask because I know that was part of the issue that led us to today's discussion on this very topic.

MR. AMIN: Thanks. So let's do

another round if there is anyone else who has comments about that. Karen, it sounds like you have a few and then Bill.

DR. PACE: I just want to say in terms of it really depends on what your analysis shows. If you indeed show that you get comparable results, which is one of our criteria, then it can be in the one measures. If that doesn't bear out, then we are to the situation where you can't say that you can just apply any risk. So that has always been part of our process.

So to give you a specific answer, really depends on what your results show. And so I mean we can have more discussion about that but I think within our current criterion process, that is how it is set up. But we can have other comments and discussion about that.

DR. WEINTRAUB: Well the problems go on and on. I think I will just mention three more. Well broad populations tend to offer better metrics than narrow populations.

So an example of this is in developing models to look at mortality after intervention of the coronaries we look at models that are in stable patients and models that are in patients in shock. And if you just at the model in stable patients, you get sort of mediocre metrics. But if you include the population in shock in that model, all of a sudden the model looks great and your summary statistics go through the roof. But it doesn't answer a question that you really want to answer because you already know there is a big difference between the patients in shock and the patients how aren't in shock. So you really have to be careful about looking at the question you want answered.

The other thing about the question you want answered is when do you include procedural details and when do you include complications? it is not always wrong to include complications. It depends on the question you are wanting to ask. Because if

you want to know what the drivers of costs are in you institution, you will find out that if you don't include complications, it all seems to be length of stay. But if you do include complications, all of a sudden complications become a big driver.

Length of stay, do you include length of stay? Some people say never include length of stay because it undermines all of your other models. But if your model is out to say what is the very best model I can use to understand all the drivers across my institutions, maybe you should include length of stay.

So very complicated. One more final one. Composite endpoints and how they handle that. I just went through this in a paper we are developing looking at long-term outcomes in stable ischemic heart disease. Should you have a model that includes just death or should it include death plus myocardial infarction? And we are absolutely

couldn't decide. There were drivers towards one and drivers towards the other. Or do you do both? But the problem of including both is then your paper becomes unreadable and people say what are you really trying to say.

So you never get to the point where you can sort of resolve all these. The problems in risk model development go on and on and on. And the best we can do is just be aware of what a complex environment this is and all the mine fields.

MR. AMIN: Okay. So we will go Karen, Andy, and then Dolores.

But quick point, on the issue that was raised by HealthPartners, I want to specifically mention that if you believe that there needs to be some change in this criteria to allow what we are talking about here, I think that is what we need, outside of having the actual data to show the comparability, that is really the question. That is really what we are trying to get out of this

conversation.

And again, so let's sort of keep to that. So Karen?

DR. PACE: I just want to respond to your point about what you include in the risk model and it depends on your question. Certainly I think that is where we need to kind of think about what we are doing with risk adjustment for performance measurement versus understanding the drivers. So certainly if you wanted to understand the impact of length of stay or complications, you would want to model that. But that is not what we are trying to do with risk adjustment when we are comparing performance.

And that is often why we don't see as big of C statistics or R squared when we are talking about risk adjustment to kind of level the playing field because we are not including all of the aspects about the care actually received. But good point.

MR. AMIN: Andrew?

DR. RYAN: So I would say that criteria 2b6 is very reasonable. And I like it but I am guessing that the incentives just aren't there for the developers to do this because it is hard enough to get one measure through a committee and the idea of having tweaks on a measure and getting that whole thing through is extremely difficult.

So it seems to me that NQF, if they wanted a kind of broader approach here would have to be more kind of muscular and really recommending testing for different types of risk adjustment strategies because I am guessing that developers just really don't want to do it on their own, number one.

And number two, perhaps the way that a measure that could be specified is that there is one primary risk adjustment strategy that is there and then there are some others that they say are comparable X, Y, and Z. And then a committee could say could recommend endorsement for just the primary or recommend

endorsement using all those strategies but could, in theory, not approve all the strategies but could maybe just improve the one that they feel the most comfortable about, which would have some more flexibility in the endorsement process.

MR. AMIN: Dolores?

MS. YANAGIHARA: I just want to register strong support for HealthPartners' suggestion. I think it is reasonable, given the resources required to get a measure through endorsement and the resources needed to review measures. For a measure that has already been reviewed, all the other aspects haven't changed. It is just the risk adjuster that has changed. A limited review seems very reasonable and prudent.

MR. AMIN: Tom and then Nancy.

DR. TSANG: I just have a very quick question about future harmonization processes as we talk about proprietary risk adjustment methodologies versus open source

and the impact of CMS's decision to use proprietary risk-adjustment methodologies and whether that has any relevance in this discussion.

I mean, Helen, is there a CMS policy about not using proprietary risk-adjustment methodologies?

DR. BURSTIN: I can't speak to it. There are some other folks still here from the CMS. I don't believe there are any clear -- yes, I don't think so.

There have been discussions. I'm sorry, go ahead.

DR. ROMAN: But generally speaking, CMS prefers to use open source publicly available methodologies. And we have encountered this in a number of areas, in particular, most recently with use of episode groupers and with the recommendation to develop a CMS episode grouper, which would be open source and publicly available.

DR. BURSTIN: We certainly heard

through the consensus process and also through the MAP process concerns that when a measure is picked up that has a proprietary sticker attached to it, the terms unfunded mandate have been spoken at these tables not surprisingly. So it is an issue that comes up.

MR. AMIN: Nancy.

DR. GARRETT: So I also support the proposal from HealthPartners for a flexible approach. I think the fact is that this measure is going to be used within markets. And I think the people who are doing this kind of work are sophisticated enough to know when they can compare things and when they can't. And just from a practical perspective, this is where we are with cost measurement right now. And this is the situation. We don't have the publicly available commercial grouper yet. So I think it allows us to start moving towards standardization without placing an undue

burden on the measure developers.

MR. AMIN: So I just want to make it clear in my own mind. Is the proposal that is being suggested now to in some way relax criteria 2b6 around the comparability of the various different risk adjustment models?

DR. GARRETT: Well my understanding is the proposal is that an entity could use the HealthPartners measure and use a different risk adjuster but still have that be considered an NQF-endorsed measure. Is that the proposal? Is that right, Sue?

MS. KNUDSON: Yes, we would test the other risk adjusters. We would only want to review the result of that test with the Steering Committee for discussion. Not the entire specification and go through what you have gone through the last two days with two other measures. But only those certain sections of validity and reliability. You know, like the 2a2, the reliability testing,

2b2, the validity testing, 2b4, the risk adjustment strategy. And then, of course, I think there is a relevant discussion about barriers to implementation if that is still relevant, given the cost of these risk adjusters. Because yes, they are commercial but they don't all come with the same price tag.

So those were the sections as we looked at the criteria, which is a much more precise review of what is different about the measure versus what is already endorsed, rather than rehash all the other aspects that don't change about the measure.

And we would only submit if we thought it was a compelling result to submit. We are not going to go through this if it is not showing consistent results.

DR. GARRETT: Would it create a new measure then for each risk adjuster?

MS. KNUDSON: Well, yes, theoretically. I mean in my mind. I am not

making the rules for NQF but in my mind that one might be one control so you don't have people comingling results of different risk adjusters. But I would let NQF field that.

MS. YANAGIHARA: I meant to say this. In my mind because it is the same core measure, it would make sense to me to have the measure number with A, B, C, D if you had different risk adjusters or something like that with a clear distinction of this is the core measure with this risk adjuster. This is the same core measure with this risk adjuster, so that you know it is the same measure but you know it is a different risk adjuster or something is different about it.

So that would make sense to me, be very clear. Because if we have different measure numbers it is like so what is the same and what is different about it. Anyway, and then that would support the limited review because it is really the same measure.

DR. BURSTIN: I mean that assumes

comparability though.

MS. YANAGIHARA: Correct.

DR. BURSTIN: That is an important assumption. I think that is what we are still grappling with is what does comparability mean in terms of what their analyses need to show to make that doable. And that is what I hope we are getting some clarity on.

DR. PACE: And if they are not comparable, then they would have to be looked at as individual measures, even if they had the same basic structure for the cost attribution, et cetera.

MR. AMIN: Jennifer.

MS. EAMES-HUFF: Just a point of clarification. When measures go through the review process and become endorsed, it is every three years that they get reviewed again. Is that correct? And as a part of that review process, is it a full review process?

So I think the request is pretty

reasonable because at some point there is going to be another full review process of the measure, of opening it up to everything. So if they wanted to have a time where you only look at a specific piece because that piece is changing, I think because you would continue the full review every three years, your maintenance piece.

DR. BURSTIN: Although we are actually only now about a year out from that, which is part of the issue of this discussion because it actually now has been a while. So does it make sense to kind of do this interim step and then, once again, within a year have them bring it all forward again? Is there a way to sort of think through really a very logical sort of step-wise approach that maybe allows us to put something out there and comment, as opposed to changing the measures and then when the measure comes back in, very clear clarity of what is required for comparability to either get to the ideal

solution, which I think is what Dolores is saying versus other approaches.

MR. AMIN: So Evan, I am going to ask you can you go to the slide looking at potential options and implications?

So I know there is a lot of information on this slide. And the side ones, there is no chance that you can actually read that. But I will walk you through what we have heard, at least today. And a lot of this depends on the information that is presented to us.

So one measure with one number, which includes multiple different risk-adjustment models, comparability of performance scores would need to be demonstrated. So in this scenario, if you are using ETGs and DxCGs or ACGs, you would have to demonstrate comparability. And it would be difficult to do that in the short-term. But even if it were done, there is some question that was raised today and in the past around

whether those results would actually be comparable. And what does comparability mean, one of the issues that Dolores raised.

The second is multiple measures with multiple different numbers. Each for each different risk adjustment model. And that assumes that the performance scores will not be comparable, meaning the risk-adjustment models are not comparable.

And that each additional measure would be evaluated against the criteria. Now that is what is being challenged here. One of our assumptions here, one of the practices at NQF is that no new measure can be created without or endorsed without evaluation of all four of the criteria. Obviously, the committee can decide to move through the criteria pretty quickly but it would still need to evaluate the criteria. That seems to be what is being raised as a question among the panel.

Specifications besides the risk

models would all need to be identical and there would still need to be some justification of endorsement of multiple competing measures, because these would be considered competing measures.

And finally, one measure with one number with multiple different risk-adjustment models and the comparability is not being demonstrated, NQF has traditionally rejected this approach for quality outcome measures. And so what would be the rationale for this being appropriate for resource use measures. And in the future, would it be possible to have performance based on these different models to be converted to a common scale.

So again we are kind of at the end of the time on discussion of this issue, but I want to just be clear on where we are ending.

DR. PACE: But it sounds like HealthPartners is saying that their initial analyses is showing that they are comparable

and so it would be one measure with different options for risk adjustment but demonstrating comparability.

MS. KNUDSON: Well I would just like to clarify that I think it goes very squarely to Dolores' point and I like the terminology use to sort out what are we talking about when we use the word comparable versus consistent results and index performance. Because if you look at the actual number we wouldn't want anyone to compare those.

And so I do think there is a couple of different nuanced positions that I have heard through the discussion that we could interpret to mean different things based on these questions that Taroon just outlined.

We, again, I would just say were compelled by the argument of this both and approach. These risk adjustment tools are tied to money moving in local markets, which is a distinct difference with the volume of

money moving based on these tools than the quality measures.

So again, it was sort of meeting markets where they are at, given where we are in the evolution of cost and resource use measures as Nancy pointed out. Yes, if you are going to do a national study, use a single tool.

So I was just simply trying to offer flexibility and I think that goes to Gene's point as well about the debate of risk adjustment at all in a national comparison, given the issues of differences in coding and practices across the country. And the fact is that what we have in local markets again, to make that local market argument, our providers are local to that market and you don't see that same level of disparity that we see when we take a national view. So I am appealing for practicality and driving an affordability agenda.

And I guess my last comment would

be with regard to the measures coming up for re-review, to me, as a measure steward, the burden is no different. If I have to go through with re-review with now six measures, I can't do it. I can do one measure with the entire spec and I can review results in the relevant sections that make a difference, based on the different risk adjusters. but it just doesn't seem very efficient to run through the same process, fill out the same application for all of that and use the Steering Committee time.

So yes, I kind of have this view of where we want to go vision-wise with what it is on the ground trying to get this done with meeting the NQF requirements and trying to advance the agenda.

So I am sorry if that is still not absolutely clear but I do hear nuanced differences in the comments.

DR. PACE: I think one of the things that we may need to do, it is hard to

kind of continue to talk about this in the abstract, so we probably need to get down to what analyses you have actually done and what they show. And perhaps that would help us figure out the best forward whether it really can be one measure.

But if the Steering Committee has some suggestions on this comparability issue, as I said our initial thinking of it was not so much that you get a one-to-one, they came out with a thousand here and a thousand on this one but that they would be basically in the same position in ranking, so it is the consistency is I think the main issue.

Certainly, if you have some additional thoughts about that but I think we probably may need to get down to some more specifics and perhaps a smaller subgroup to help us sort it out.

And one question. This issue about markets using the same risk adjuster or model, I guess one of the things that we would

be interested in knowing is will providers in those markets be getting -- so is it possible that providers are going to get one health plan that is using a particular proprietary model versus another health plan? So a hospital may get two different results.

So when people say markets are using one risk adjuster, I am not sure how accurate that is. So I guess that is what I am trying to understand.

DR. NELSON: It is what you suspect, that within one geographic area, one health provider may get many different risk-adjusted results.

DR. GARRETT: But it is usually within the same payer. They would be using the same approach across all of their providers.

DR. NELSON: Right. Payer is consistent.

DR. GARRETT: So kind of the current -- yes.

MR. AMIN: So from a provider perspective, I mean this whole issue about harmonization and competing measure from a consumer purchaser and provider standpoint is that if you are introducing multiple measures with multiple different risk-adjustment models, you are still introducing a significant amount of burden, potentially.

So it is a question, right? Okay.

MS. CLARK: Yes, I was just going to say I like the idea of the whole comparability and if you are using one risk-adjustment methodology versus another, as long as the rankings are the same and the people come out in the same types of groupings. But if the actual value is different, I mean, this has come up before and I am thinking of an example in cardiology.

For example, they have a EuroSCORE and Society for Thoracic Surgery score for looking at risk adjustments for mortality. These are two very different methodologies.

One is proprietary, another is not. But people kind of know that if you use a EuroSCORE, you are going to get like the mortality rate as through the STS score.

And so if there would be some way to come up with an adjustment that somebody would know if they were using ACGs versus DCGs versus ETGs, you know you have some type of way of comparing them. That would be --

DR. PACE: I think that is a good point and it gets back to Dolores' question so that even if the provider would have the same ranking, if you try to use these across for cross-comparisons, we are still in a problem.

MS. YANAGIHARA: A quick comment on the market question. I mean I think there are certain markets that are coalescing around a particular adjusters. So in California, we are using DxCG across the state, at least for the program that IHA is responsible for. It doesn't mean that health plans have other adjusters for other programs. They probably

still do but at least for the statewide P4P program, we are using the same thing across the state. I think Minnesota has just come to agreement on what they are going to use.

So I think there are a lot more instances and I know that a lot of the states in the state innovation models on the CMI funding are very interested in total cost of care and would be very interested in having -- using an endorsed measure.

And so to the extent that a public grouper is available and can be tested and that could be used, that would be great. But short of that, if there are options, I think that the different states would be wanting to use different things, depending on what they are already using.

MR. AMIN: Okay. So thank you, everybody. This has been a very robust conversation and thank you to our colleagues, developers and those on the phone.

So we will sort of take this

offline, think about this, digest it, and provide it back to the CSAC in terms of a digestible form in terms of where we want to go.

We have an attribution discussion and a sort of a forward-looking discussion in terms of how to really move toward efficiency. We are supposed to break at 2:00 and we have lunch. And we want to be respectful of people just wanting to take a little bit of a break. This has been a very intense morning and it is already the afternoon.

So again, let's break for ten minutes. I know that is not much of a break but we want to at least try to have half an hour for the last discussion. So I hope that is okay.

(Whereupon, at 1:10 p.m., a lunch recess was taken.)

A-F-T-E-R-N-O-O-N S-E-S-S-I-O-N

(1:25 p.m.)

MR. AMIN: Okay, so again for a little bit of context, this next discussion as many of you who were on the first resource use Steering Committee remember, there are components of the measure submission that were allowed to be submitted as guidelines or specifications. And specifically what that means is that the measure developers could elect to provide guidelines or recommendations on how the measure specific portion of the measure construct could be applied in various different context or they could provide specifications.

One particular area that that still exists is in the attribution, where currently there is this allowing of guidelines or specifications.

So specifically NQF, again going back to this, NQF endorses national standards for performance measures and that includes both quality and resource use measures that

are intended for both accountability and public improvement.

Actually, let me turn it over to Karen. Actually this is Karen's section.

MS. WILBON: I will just make one clarification. When we say guidelines and specification, specifications are what is baked into the measure and is standardized. So we endorse what is specified.

The guidelines would be used by an implementer or user who may need -- may not have any idea of what attribution model they might want to use. And so the guideline was there for suggestions for the user but it wasn't necessarily part of what was endorsed. So to say that a user who was using the endorsed measure must use this attribution model or must use some other guideline that was provided by the developer.

In addition to the attribution and you might have seen it in the appendix, there is other pieces of information that we also

allow developers to submit as guidance. So how they would suggest that benchmarking be done, how peer grouping would be performed using the measure, all of these things we have been categorizing and are reporting more as a reporting function, as opposed to actually being baked into the -- I used the word baked but kind of specified and baked into the measure.

So I just wanted to clarify that so that everyone was on the same page.

MR. AMIN: So again, since the fact that NQF endorses performance measures for national comparisons as national standards, the question really becomes in order to make useful comparisons about conclusions of performance, especially relative performance at a national level, all entities need to be measured in the exact same way. And this again goes across quality and resource use measures.

And specifically in quality

measures the attribution is part of the measure specifications and is required for submission. So to the extent possible, NQF criteria should apply for all types of measures with only a minimum number of exceptions that are absolutely needed for specific types of measures.

So the issue specifically that we would like to discuss is what would be the rationale that resource use measures should be handled differently? And I will turn it over to Karen to maybe provide a little bit more.

DR. PACE: And I think this is generally the same discussion we just had about risk adjustment models. And the bottom line is if you get different results using different attribution rules, then it is hard to say that it is the same measure because it is not standardized. And how could you compare?

And I think there is definitely more attribution questions and components in

a resource use measure. But as Taroon said, for quality performance measures, how people are assigned to whatever level of analysis it is, whatever entity is being measured is part of the measure. And once you start having a lot of flexibility, the question is what are we endorsing as a national standard.

MR. AMIN: David.

DR. PENSON: So, to start the discussion, I think this may be a very short discussion because I got on the assumption that this was required. And to say that it is a guideline, I would say that you are better off having a guideline for the quality measures than you are for the resource use measures when it comes to attribution because attribution is so critical here, how you categorize the different resource use categories and who does what, and particularly the way these are going to end up being used, I can't really think of a good reason just to have it as a guideline. I am sure others may

disagree but I don't think many will. So I don't know.

MR. AMIN: Cheryl?

DR. DAMBERG: I would agree with that comment. We have done various analyses looking at different attribution rules and you definitely get different results. So this just seems like I'm not sure why we are dealing with this. I think it definitely should be part of the criteria, the requirements.

MR. AMIN: So maybe as we continue going forward before we get to -- maybe the straw person is that we move the attribution components and the reporting components, meaning the peer grouping and sample size requirements components of the measure submission to be specifications as part of the measure. So Lina -- okay.

And if there is disagreements with that, let's have that discussion, since that seems to be the general tenor of the

conversation. It is only two people but let's start. Jack?

DR. NEEDLEMAN: As somebody who just voted no on a measure over the attribution rule, --

(Laughter.)

DR. NEEDLEMAN: -- I actually like the idea that the attribution is a guideline because I don't think we know the right way to do attribution yet. And what I can see is it is important to figure out which costs you are going to count, how you are going to deal with the outliers. So you standardize what costs are being counted and what costs are being aggregated.

But in a lot of the systems in which these measures are being used, there will be either negotiation between the insurers, typically, who are the payers who are actually producing the data and the providers about the way the attribution should be done. And that is subject to negotiation

that makes the rule acceptable. And given the uncertainties of that attribution I am willing to see those negotiations go on.

And because the way the data is going to be used once it gets to the providers in terms of figuring out how we manage these costs, which is ultimately what we are talking about, not who is to blame but how we manage, allowing for some flexibility in the way the attribution is done gives the provider group some opportunity to reflect on the management problem, rather than the attribution problem.

So for those reasons I think that there is a certain value in allowing the providers who are the recipients of this stuff and the payers who are the compilers of the measure to negotiate what the attribution structure is going to look like within some broad guidelines.

MR. AMIN: Lina, Dolores, and then Larry.

DR. WALKER: I have a clarifying

question. I guess I don't fully understand the difference between requirement versus specification. I was just listening to Jack and you were asking for flexibility in the way the attribution is done. And so how is that precluded if it is a specification?

DR. PACE: Well, basically what it means is that we endorse the measure as specified and so I guess I am not sure what people mean by guideline because guideline tends to mean you can do it multiple ways. This is one way but if you choose to do it another way, you could versus if it is really part of the specifications, that is what NQF has endorsed the measure as specified.

DR. WALKER: Can I ask a follow-up question? In that case, I mean I guess I would ask Jack this question. We evaluated the measure and, in fact, you voted the way you did because of the attributions. Why would you want to provide more flexibility?

I mean I guess we have to evaluate

it in the context in which they say it would be measured or computed. And that hinges on the attribution. So I don't even know how you could evaluate it if you give them flexibility.

MS. YANAGIHARA: This is a really complicated thing because it really depends on the use case. How you attribute really depends on what you are trying to do with that measure.

So for example if you are trying to encourage medical homes, you are going to want to attribute things to one group or provider or whatever, probably group, to really try to drive that that is where you are responsible for these people. Whereas, if it is for a different purpose, you might attribute it differently.

So I don't think it is -- and I don't know that you want different measures that are exactly the same with different attribution methods either. So it is just

really hard because it is much, especially resources there is so much tied in to what are you trying to accomplish with the measure.

And so I don't know without some flexibility how you can -- I don't know. I think you are just asking for more people not using NQF-endorsed measure tied into a particular attribution measure doesn't really align with your use.

MR. AMIN: Joe?

DR. STEPHANSKY: As an example, I am going to go back to Blue Cross/Blue Shield of Michigan again, where the attribution methodology was debated heavily for over two years before they finally came up with something that was acceptable to both physician groups and the hospitals as to how patients were linked together across provider groups. Two years.

But now, everybody is at least adequately happy to move forward. If we would have had to start with a specific attribution

methodology as in what we were looking at in the last measure, we would never be there. People would have simply rejected it.

When we go into the commercial market, we have to recognize that these are negotiated contracts and measures are now going to be part of those negotiated contracts.

MR. AMIN: David.

DR. REDFEARN: I would like to see some more flexibility from possibly sort of a selfish personal interest. I would like to see some attribution methodology that is something other than that stupid Dartmouth visit count method, which seems to have taken over the world. And the reason I say that is in California when we first started our ACO pilots in California, I was asked to do some analytics. And because I do so much episode work, I thought I would take a look at our episode data and see if I could use the episode data, which you assign an episode

responsibility for an episode of care to a physician and then that links the physician back to the patient whether I could use the episode data to build an attribution model driven out of the episodes of care that the physicians were managing. And I developed that methodology and reported it to our network people and to the medical groups that we were dealing with at that time, which was Healthcare Partners in Monarch. And to my amazement, people liked the methodology.

So the initial ACO pilot for Blue Cross of California was based on an episode of care attribution methodology. I did some comparisons with the Dartmouth methodology and I think I demonstrated that the episode methodology of assignment was superior, was really linking patients to doctors that they were really treating and not just had one or another visit.

And then but to my dismay, the corporate ACO pilot decided to adopt the

Dartmouth methodology and that has now even disappeared in California. So all the work I did is just gone.

So I guess what I am saying is if you have some flexibility to start looking at some potentially superior methods of doing the attribution, that is not a bad thing. I think that is a good thing. But then like I said, I have a kind of personal interest in this because I have been through this kind of process and in fact, when I go to the management and said well I think the episode data is probably superior and gives you some value. And it is like oh, yes, but it is too hard to do. It is too complicated plus there are these standards that we have to use the standards.

I even had some discussions with the Dartmouth folks that were working on the method and they even sort of admitted that there were some limitations to that kind of methodology but said we are working on

improvements but it is what it is now and that is what we are doing.

So I like that additional flexibility.

MR. AMIN: Lisa?

DR. LATTS: So I think that actually this whole discussion is an artifact of our times because five to ten years ago, nobody ever talked about attribution. So it is really this whole discussion about the methodology is a reflection of some reluctance by whoever it is being attributed to be held to this standard.

I think that within several years there will be a clear winner but right now there isn't, which is why we are even having this discussion.

I'm okay with the flexibility as long as NQF or whoever is endorsing the measure comes up with a set of principles that the flexibility has to adhere to. So I think that if we do allow flexibility in the

methodology, we then have to say you have to meet these five principles to insure consistency. Otherwise, why are even spending time approving measures because the attribution could so screw things up that it might not even be reproducible.

MR. AMIN: Daniel.

MR. WOLFSON: So I think attribution really has to do with accountability. Right? We are trying to measure who should be accountable. And I know my people on my right and my left won't want to hear this but I do think that the accountability models will change rapidly over time as a source of information goes away from health plans and goes back to the delivery systems. And we are going to see a much different way of thinking about attribution accountability in the future. And where the source of information will be, the rich clinical information which we really need to have, not claims data, will switch.

And so we need to be flexible about those attribution models because the attribution models today will be obsolete. I don't know when but they will be obsolete. The shift in power and influence, I know this is self-serving because now I work for the physicians and not health plans. But I think it is true. I think it is changing because of the EMR. And if we are not flexible and aware of those changes, we will be in an old chassis.

DR. PACE: I think the question is, obviously as we have talked about, the measures are only endorsed for three years and then have to come back for endorsement maintenance. So we are not talking about that once we endorse a measure and an attribution rule that would go with it that it would never change. That is not the model of NQF endorsement in the first place. The question is within that three-year period, if you have endorsed a measure and you all are saying that

the attribution is so key to that measure, how can you even, what does validity of the measure even mean if someone can use a different attribution rule?

I'm just trying to understand what it is you think you want NQF to be endorsing if it is not -- if attribution rule is not part of it.

MR. AMIN: Okay, Nancy and then we will take some -- oh.

MR. WOLFSON: You can't have measurements without attributions.

DR. PACE: Exactly. And it is part of the specifications of all of our other measures.

MR. WOLFSON: Yes.

DR. PACE: And the question here is whether attribution should be kind of outside of the specifications and allow multiple ways or people can select their own attribution model.

MR. WOLFSON: I can't comprehend

how it could be outside. I just --

MR. AMIN: Okay. So just for clarification, that would actually be a change, a shift from where we are now and alignment with where we are in quality. And that is sort of the straw person that we are testing here, to make sure that we are all in agreement that we are okay with moving in that direction.

So we will go to Nancy and then comments on the phone, then comments in the room, and then we will go back around to Bill, if that is okay.

DR. GARRETT: So I agree with the idea of having attribution be a standard part of the measure. One of these uses for these total cost of care measures is going to be and already is helping to shift, starting to shift accountability, financial accountability for population to providers. And as that happens for providers to have slightly different or even very different ways of defining that

population with all the payers they work with, it gets really complex quickly.

And I think that this is just an important role for standard setting to be able to start to move toward a national standard. I think it is really an important thing. And we won't do it right at first but it will have to evolve over time, I think, to get better and better in terms of what are the best ways.

And unlike the previous discussion, we don't have the same complexity with risk adjustment where there is a need for proprietary method. I mean attribution is something we can write down on a piece of paper and have it be completely open source and we can all come to an agreement about the standard.

MR. AMIN: Okay, Operator, are there any comments on the phone?

OPERATOR: At this time, to make a comment, please press *1. There are no comments or questions.

MR. AMIN: Thank you. Are there any comments in the room?

MS. KNUDSON: I guess just reflecting on the comments that the Steering Committee made, I point out three things. First of all, I think we are moving to a model of shared accountability with greater accountability with those risks with the providers. But I don't think the insurance companies will not have any or health plans won't have any accountability. I don't think it is a full shift. So I think I would offer that I would view it as shared accountability.

And as it relates to working for an organization that has the benefit of understanding both EMR data, as well as administrative claim data, you really need them both to understand resource use and total cost of care.

The EMR is not a replacement. Because when we refer our patients to our partners in the community or they get

hospitalized outside of our own system, that is where we use our administrative claim data to understand those pieces. And that is not in our EMR. It might be there with some coordination of care notes but in a real easily retrievable way. And those are the realities of this work on the ground.

I think the other comment I would make is the attribution methods. It would seem to me that when we have a single payer like the measures we just listened to from CMS as a single national payer is sort of a different bailiwick to have consistency then across different markets with commercial payers. And this is where I would say just like based on the experience we have had in Minnesota, we spent probably nine months to the earlier comment vetting the different attribution approaches and landed on one across the different payers. But I am not convinced what is going to work in Minnesota will work in every other local market for

commercial payers.

So I think this again is a topic that is not black and white. There is a lot of gray space and particularly in the commercial market. And so if we really want to advance these, we should not be too overly prescriptive in the commercial market with a one size fits all across the entire country. I think there needs to be some flexibility for different market places.

And it will evolve, just as I mentioned before, that I don't view our Minnesota work as static. I think it is going to evolve as we redesign and improve care delivery in a way that brings in other sort of virtual and alternatives to face-to-face care where we are not counting visits. We have to be more creative than that.

So with these issues, our work will never be done. I guess is the main point and we need to remain flexible.

MR. AMIN: Thanks, Sue. Bill?

Oh, was there any more public comment? I'm sorry. Okay, no.

DR. WEINTRAUB: So flexible but within limits. And every time in a modeling exercise of any kind I hear flexibility, I always get worried because it opens the door to multiplicity.

And so a certain amount of flexibility but specified in advance so that people don't sit off and say well they just will do whatever they want to until they find the answer that they like. That is just not valid when people do that.

MR. AMIN: Jennifer.

MS. EAMES-HUFF: I think this is a reflection somewhat repetitive of what people are saying of where we are in a point in time. I don't think we have enough evidence to know what is the best attribution method or which one is appropriate for which use. So it is hard to create a single standard without having a sense of having enough evidence

around it.

Yet, this is a standard setting organization and I think ultimately we would like to move to some consistency across the market. So it seems like there needs to be some way that maybe a role NQF could help start driving standardization and consistency while we are still in the phase of evidence gathering and trying to understand. I think that is perhaps where some of the guidelines or principles could come into play in terms of what would be guidelines or principles that the attribution methods would be sort of assessed by.

Because I think the other thing, I will just raise as another issue, is if the measures don't end up in the same committee, you get the variance across committees as well, like it becomes a different group reviewing it. So there needs to be some level of consistency across groups as well over time. And I think that is where the

principles come up as well.

MR. AMIN: Lina and then Jack.

DR. WALKER: I have a clarifying question. So is it the case that when NQF evaluates a measure that the committee would look for best in class on all the different components of the measure and also looking for broad applicability?

So I mean I hear some of the comments about like there needs to be flexibility for local regions or some even using the term standardization. But it seems to me that if the goal is to evaluate the merits of the measure, you want to -- the developer should be looking at the best possible approaches to achieving that particular stated goal. And to say that there could be different types of attribution approaches, which presumably could result in different outcomes for that measure, makes it a little bit harder, I think, for the committee to evaluate that particular measure

for endorsement.

DR. NEEDLEMAN: I think it is useful in thinking through the issue of getting it right. Because I am not sure that there is always a best in show but we are talking about trying to get it right. That when we look at how measures get developed, they get developed in two very different contexts and we have heard that referenced today.

In some cases, substantial elements of a measure are developed in negotiation with a lot of iterations from experience of folks who are going to have to use the measure.

So we heard Joe talk about the negotiations in Michigan with providers about attribution. And there are lessons that can be learned from that because I suspect there were a lot of different things on the table. Some data, but perhaps not always data being analyzed there but some data saying well what

is the implication of doing it this way versus that way. And you get certain experience there. And I heard I think Nancy say a similar sort of thing happening in Minnesota.

Others of these measures, and I can say this because I have developed some, are developed with the data in the privacy of your own room. And there are some strengths to doing that. The validity, reliability kind of testing can take place and take place pretty quickly. But they are not necessarily tempered by the experience of how does this have to be adapted in practice. And I have no idea how the Dartmouth people actually did their attribution model. But if you asked me to place a bet today, I would be betting on in the privacy of our own room, which is fine. But we have got a process where we have got to think about getting it right and learning by doing. And the question is, at what point does the endorsement have to narrow the range of choices in getting it right?

And there are some areas where I think it is clear just looking at whatever is coming in that something was done wrong and should be reversed. There are other cases where it is not clear what the right answer is. I think we saw that with risk-adjustment. I think we see that to some extent with some of the different attribution models we have seen walk in here in both Phase I and Phase II of the resource use measure.

So what we need is to, I think, allow some flexibility. But even more important, we need to think about how we are going to learn.

So if somebody were bringing in a measure with attribution based upon the Michigan negotiations or the Minnesota negotiations, I would want to hear how did you get there. Because how you got there is an important part of understanding why this attribution method, why you think this attribution method works. What are the

alternatives? What happened when you tested them?

We heard the CMS, the Mathematica people say we tested a lot of different attribution models. And basically, all the patients wound up in the same groups most of the time, almost all of the time, actually I think was about the characterization. And that is powerful and that is working in the privacy of your own room but it is a powerful test. We tried it a lot of different ways and we come up with very similar results and then we picked this one because within this range of things, it doesn't make any difference.

But I think we need to hear those stories in order to make decisions about whether this is good enough to endorse either as a standard or as a guideline or to suggest how much range within guidelines should be allowed when we are not quite sure what the exact right answer is.

MR. AMIN: Gene and then Karen.

DR. NELSON: Jack always makes great points. That attribution word can't happen at the other end of the hall, so I am not quite sure if it was highly collaborative or in their own offices.

(Laughter.)

DR. NELSON: I will check into that. We think we are good at partnering with real places but we might not be.

But the issue of attribution, just a general thought is that as we try to get alignment across different kinds of performance measures, be they outcomes or costs or resource use or technical quality for patient experience, there is the different levels of performance. So it might be a provider. It might be a hospital. It might be a nursing home. It might be an ACO. It might be a medical home.

So for me, I think one of the issues is to make sure that the attribution is consistent at different levels of

responsibility so that if it is a provider level measure, an individual doctor level measure, there should be an attribution method across the different quality measures. If it is an ACO, there should be an attribution method across ACOs. If it is a hospital, it is simpler. But I think that is one point is that the attribution methods should be similar across measures for different levels of the health system and that we should be trying to align the attribution method across all the quality measures.

MR. AMIN: Karen?

DR. PACE: Yes, I just wanted to clarify that I don't think that we are proposing that we are ready to pick one attribution method that applies to every single measure. I think we are just trying to sort out whether, in our terminology, if it is a guideline, that means it is kind of voluntary. And it sounds like when you all have been evaluating measures, you have been

evaluating whether it is called a guideline or specification for the attribution method.

And so I guess my question is if you endorse a measure and whether the attribution method was called a specification or a guideline, would you expect that that measure could be used in somebody say well the attribution method was a guideline and I have decided I am just going to use a totally different attribution method that I have made up. Would that still be what you intended by endorsing a performance measure?

So I guess maybe we need to get to what we are talking about with the guideline and what that implies for endorsement versus -- you know, maybe we are just talking about words with no distinction or maybe we are really talking about something very different.

DR. DAMBERG: Maybe I was naive coming into this process but I guess I thought it was part of the specifications. So it is kind of interesting to hear at the end of the

meeting it wasn't.

And I guess then the question is, how would that have changed my evaluation of what I was looking at over the course of the past two days? And I mean I totally get the need for flexibility here. And I definitely think it is worth considering more. Are there certain specific applications where you would want to call out, shouldn't be used or attributed in these ways. So maybe start from that side, rather than sort of saying it can be applied in like a thousand flowers bloom approach. So call out where it shouldn't be done.

I don't know. I sort of feel like there needs to be a few bounds put around it. Because I am not sure we would have all agreed that the total cost of care measure could be applied in 15 different ways.

So I think it is just hard to evaluate the measure without having some context in which you are applying it.

MR. AMIN: It sounds like we got a pretty robust conversation around this topic with various different perspectives but generally coalescing around where you were going, Cheryl.

Tom. Let's kind of end it with Tom and then I am going to turn it back to Ashlie to talk about our gaps discussion for our last hour.

DR. TSANG: I don't know if I can kind of tweak Jack's comment about using the word flexibility as much as more of the parameters and the descriptors of the methodology itself.

So in my mind, in the future I would want to understand like the transparency of the methodology or the robustness and the reproducibility of that methodology itself, not so much as is it right or wrong but rather the parameters of how the organization, how the measure steward and the developer came up with that methodology.

And so with the CMS measure, like we understand. They explained it and we understand. And you may or may not agree with it but I think we understand the flaws and the characteristics and the strengths of that methodology.

And so that is what I would take home as the future evaluation process of attribution. Does that make sense?

MR. AMIN: Yes. Okay, so I will turn it over to Ashlie for our final piece of today's discussion.

MS. WILBON: Thanks. So I know we are kind of dropping like flies. So we are going to -- this last discussion is really more of a kind of 50,000-foot forward thinking kind of providing us some insight on where we go next in this particular measurement arena.

I do want to have one order of logistics or what have you. Lindsey passed around I think to some of you whose laptops probably weren't open, a piece of paper with

some survey questions on it. We really want to hear you feedback on how your experience has been on the committee thus far. I think she also sent out a SurveyMonkey link via email. So if you are not able to do it on paper before you go, please on your way home or something while it is fresh in your mind, submit some feedback to us. We really appreciate that. So I just wanted to say that before we lose any more people.

So this last discussion is something that we generally do with our committees to figure out kind of where the next steps are in a particular measurement area. Obviously, this is only our second project around cost and resource use. So in terms of the number of endorsed measures that we have compared to the number of endorsed quality measures is obviously, a lot less. And so if we are looking at gaps, obviously, there is going to be a huge gap because we just don't have a lot of endorsed measures and

cost measures.

And as you have gone through the process these last few days, you can see how much time it takes not only for reviewers but for developers to develop measures. But we would really like to hear your input on kind of the types of cost and resource use measures that we really should be looking for. We are really trying to take on a role with working with developers earlier on as Taroon mentioned, to really give them some insight and some feedback on where the development is really needed and where they should be focusing their money and time.

So insights on that, obviously we have talked about a lot of different issues. We don't necessarily don't need to go back to risk adjustment and attribution. But there are some other issues around resource use measure and our cost measurement broadly within the context of maybe where we are going with the healthcare system that you may want

to identify. We would be interested in hearing those.

In addition to doing endorsement projects, we do also convene experts around specific topics, like we were talking about earlier convening people around to talk about risk adjustment and the impact on quality and resource use measures. So those topics are also of interest.

And then kind of this forward thinking thought, your thoughts in terms of where we go with the sufficiency measurement issue. We have been, I think, since 2009 since we did our first resource use project, have really been looking forward and trying to figure out how we actually make this linkage between cost and quality measures. What does that actually look like and where do we start? Is it just sitting down and figuring out which quality measures, how many quality measures you actually link with a cost measure? Is it just one? How do you do that across an

episode? How to make a patient-centered. What things need to align? Do risk-adjustment methods for the outcome measure that you link with it need to be the same risk adjustment that is used in the cost measure. You know there are so many different issues there but any insights you have on that would be useful.

I know I am throwing a lot oat you but we have only got 30 minutes, so just throw out what you have got before you have to leave.

And then a much bigger issue, which I am sure we won't be able to get to all of these but suspect, at some point in time, that we will be evaluating groupers or potentially a grouper. And even episode-based measures and any thoughts in particular on how this evaluation process, we have looked at total cost kind of per capita measures, and what specific ideas you have around how the evaluation criteria applies or the discussions that we have, how that might apply to

different types of cost and resource use measures beyond just the per capita total costs measure. I guess we did have the hospital episode-based measure. But kind of the more condition-specific focuses I think is where I was going with that.

So four really big buckets. Again, we are just looking --

DR. NEEDLEMAN: We are here for another two days, right?

(Laughter.)

MS. WILBON: Yes, we could probably have two-day meetings around each of these things. But again, this is just kind of getting your insights on some of the bigger issues that we are kind of grappling with and where we go on next steps.

So if anyone has any thoughts, we are open to hearing them.

MR. AMIN: Daniel?

MR. WOLFSON: I have to say this before the meeting is over. And I know this

is a MAP issue. But to think that we are going to only look at existing cost and not appropriateness misses the mark. Because everything you are measuring, you think it is appropriate. We know that the measure of appropriateness is about 30 percent is not appropriate, is not necessary and is waste.

So it baffles my mind that we are focused on existing costs and assuming that every admission is appropriate. And we know that is not true.

So I had to say that before the meeting -- I know that is a MAP issue but I waited until the end to say that.

DR. BURSTIN: It is actually not just a MAP issue. I mean, broadly we bring appropriateness in our measures to every single clinical committee at NQF. We have actually got a couple dozen of them now. We are trying to work with the specialty side to get additional ones in. It just isn't appropriate for this particular committee of

cost and resource use.

MR. WOLFSON: Okay.

DR. BURSTIN: It is usually required in the clinical ones.

MR. WOLFSON: I knew it wasn't appropriate to say. I waited until the end.

DR. BURSTIN: No, you said it.

MR. WOLFSON: I know it wasn't Helen. I bit my tongue for two days.

And then too, I have heard it said that we have to wait for outcome measures, I don't know if that was true, of quality to bake into an relate to the cost. And that worries me because the outcome measures are so underdeveloped that can we use intermediate outcomes. I mean are we really talking about outcome measures? Because I worry that will impede the progress of cost measures, if we are waiting for the ultimate outcome measure to correspond to it.

And then if we get into process or any immediate process, it really baffles the

mind about what would be the associated process measures to go with the cost. It would maybe -- mortality is pretty crude.

So I think that is a real huge challenge but I worry about waiting for the ultimate outcome measure.

MS. WILBON: Yes, I will just comment on that. I think you are probably getting that from one of the MAP input slides that we had, where they were more focused on pairing or linking cost measures that would be endorsed with outcome measures. I think that was probably their preference to have outcome measures but would it be the end all/be all that they wouldn't move forward with recommending other quality measures that weren't outcome measures.

And if you really kind of look across the episode, an episode of care for a patient, there is various areas across that patient's interaction with different healthcare systems and providers where process

measures would be appropriate and obviously outcome measures as well. So it doesn't exclusively mean outcome measures. But I think there is a preference to have at least some outcome measures in that picture.

MR. WOLFSON: Helen, how can I get my hands on the appropriate measures that you guys are reviewing?

DR. BURSTIN: We have a list we have been generating. I can send them along. We need many more but we have got a good number to start.

MR. AMIN: Bill, Jack and Jennifer.

DR. WEINTRAUB: So anyway, it has been a great couple of days. I always say that I learn much more than I could possibly contribute. I think the process is fundamentally a good and sound one.

In terms of what measures, I mean these two measures, these last few days, these were big ticket items. And there are big

ticket issues before us. And I think this is appropriately where we should be, look for the things that really are critical to our healthcare system.

You know I was on the last one, too, with cost and these were bigger. As you and I had that discussion, these are bigger and more important than the last go round. And probably appropriately so because we were sort of treading water or learning how to do this the last time, dipping our toes into the water I should say.

And clearly a lot of time is needed for discussion. The last time we didn't have enough time. There were too many things we tried to do in too short a period of time. This is really good that we could sort of really dive in and everybody had time to say what they think.

I just want to comment a little bit on how we are going to relate clinical outcomes measures to cost measures. No easy

task. I have been involved in cost effectiveness analysis, as some of you are probably aware for 20 years. And that kind of methodology doesn't really apply here. At least I don't think so. But I think a lot of thought needs to go in on how we are going to make these work together. We are probably not there yet.

MR. AMIN: Jack?

DR. NEEDLEMAN: A couple of issues that strike me is important is one based on the past experience both in phase 1 and in this phase. We talked about attribution to providers today a lot and that is where we finished. But we have also got the question of what the scope of the costs are. And that has come up when somebody gets hit by a bus 500 miles from where they live or 1,000 miles from where they live, do those costs get lumped in? Is there any way to differentiate that? So you have got that issue with an all-cost measure.

When we were looking at specific diseases on the last go round, it definitely came up which costs. We had some measures which were patient, patients with diabetes but then all costs. But we also had some measures which we were trying to get at the episode issue. And there are all kinds of issues in figuring out what gets counted in the episode. So we saw that a little bit with the first measure today.

So that is going to be an ongoing issue and one in which we are going to learn through the struggle about how to think about those. But we need to make sure that -- so that is going to be one of the critical issues.

We ought to be thinking about some not all-patient all-cost measures but I think that I would like to see us revisit things like diabetes and some of the cardiac patients, and some of the other patients where we know that they have got high, potentially

high costs. And we want to look at what the variations in the costs are and what the variations in the outcomes are.

I think in terms of the risk-adjustment models, as we look at this all-cost thing, an issue I raised almost offhandedly the other day, earlier today or yesterday, I lost track, was thinking about how well the risk adjusters do in predicting non-discretionary use of specific services.

So how do we think about the hospice costs being in here? How do we think about the readmissions? How do we think about the SNF costs? If you have had a hip replacement, you are going to be in a SNF for a few days.

So thinking about how well the risk adjusters are doing it at differentiating the discretionary costs and perhaps the excess costs, allowing us to pull out the nondiscretionary elements and think about that are going to be one of the directions we are

going to be asking people who are doing risk adjustment to be moving into or at least analyzing as they get these more global measures of costs and resources.

In terms of the outcome -- in terms of measuring the other side of the value thing, which is the outcomes, lots of issues with process measures, including their completeness. Lots of issues with outcome measures, including how much they are attributable to what is going to the actual care. And those need to be resolved.

I actually like the CMS display of the grid, showing the performance on the measures that were there and the performances on costs. Because I think that is a good way of not getting a single number but looking at the way things distribute. The question there is were the process measures that were included in where you got on that grid in terms of value sufficient, adequate, complete enough to be a full measure of the quality of

the care that one was getting. So I think the completeness of those sets and how they get pulled together in measurement is going to be another issue that is going to come up as we begin looking at the value side of the efficiency measurement issue.

So multiple measures of outcome, multiple measures of process. How they get integrated and coordinated into a composite measure is going to be an important issue for this committee or the next committee that is dealing with this set of issues.

DR. NELSON: Jennifer, and Dick, and Tom.

MS. EAMES-HUFF: If we are truly going to get at the affordability issue, we have to go beyond just looking at resource use and standardized prices. And we need to look at the prices that are on the market.

We know that resource use only plays a portion of the variation in the market and that prices widely vary and that that is

contributing to the rising costs and the unaffordable healthcare.

And I think this becomes even more important as we are moving towards more coordinated care, which I think is the right direction, improving quality, having better coordination but it is also bringing providers together and creating market power in some instances, which we could be improving quality but we could also be improving prices at the same time. Well not improving -- I mean increasing. That is not an improvement.

MR. WOLFSON: CMS took a big step yesterday doing that for at least 100 procedures and what they cost around --

MS. EAMES-HUFF: What they charge, which charges don't really mean anything.

MR. WOLFSON: Well they showed variations.

MS. EAMES-HUFF: Yes, but they are not --

DR. NELSON: So, Dick. Dick, your

card is up. David -- sorry.

DR. REDFEARN: I just want to put a plug in for more episode-based measures. Because I think to some extent it adds some complexity but I think it simplifies attribution because you know what are you attributing to taking care of that episode. I think it also makes it a little bit easier to choose quality measures. It simplifies that.

So I think there is a lot of advantage to that. The disadvantage to you guys is you are going to have a lot more measures. And for us, that means a lot more time looking at measures. But I think there are some inherent advantages to those because they focus very nicely.

And the only other thing I would add is that just I would suggest to reach out to the vendors, and particularly to Optum to try to get them involved in this process because I know we have got some episode

measures coming up and I actually talked to Tom Lin and said are you guys going to submit and he said no. He said we had a bad experience and we are not going to do it again.

So I think it would be nice if you could reach out to them again and encourage them because I would hate to see you guys lose that. They could still choose that they don't want to do it.

MS. WILBON: We have been in contact with them and we have had some calls with them. So yes, thanks.

DR. NEEDLEMAN: We have nothing to apologize for.

DR. NELSON: Tom?

DR. REDFEARN: this is actually a reference to the Gretzky Group and the work that you guys have done.

And I still have very much faith in this eMeasure business. And it is related to Jack's suggestion about a composite measure

combining these type of resource measures, along with outcomes measures. And I see a real future in doing that with some of the outcomes data that you can gather and derive from the EHR clinical data.

And if there is a way and a mechanism of whether it is this group or whether it is calling out to measure stewards or whether you are going to incubate this process, I think there is a real opportunity to leverage and interrogate some of the clinical data and combine it with some of the patient-reported outcome data, as well as a resource measure and really create a robust platform for a composite measure.

DR. NELSON: Thank you. Taroon.

MR. AMIN: Actually, it is very much along the lines of Tom, your comments. it is actually a question. As we look to sort of the conceptual framing, just as everyone walks out -- oh, Evan, if you could go to the building block slide.

You know this has been one of the sort of conceptual framings of this Work Group, in addition to the patient-centered episode of care framework. But well actually this is sort of where I was going. So when we look at the cost and one of the things that we have talked about over and over again is how do we get to the measures of efficiency and value. And one of the hypotheses that started this group was that let's really get sound measures, sound usable measures of resource use and then we will be able to get toward measures of efficiency and value.

What are we really looking for when we were talking about efficiency measures? Do we really think, sort of Tom and Jack sort of pointed out, are we really looking for some type of composite that can actually build these two various different sort of conceptual domains together or are we may be looking for something that is not in an individual measure, yet in a way that is

reported or more like program features, meaning something that we see in sort of value-based purchasing where you have various different domains and you weight those domains in a certain way.

So what exactly is it that we are looking for and how do we know we get there?

DR. NELSON: I'm glad you asked that with six minutes left. That is an easy question.

MR. AMIN: Sorry, Gene.

DR. NELSON: So Jack and we should have a pause for public comments and then wrap up the next steps.

DR. NEEDLEMAN: Just one last comment on the journey we have been on. And this may be not our job here. But once again, all of the resource measures that we have looked at have basically been generated by claims files, by billings. And that has very important implication for actually measuring resources. There are all kinds of services

that individual groups and physicians' offices and health systems are providing, which are affecting care, which are not billed for, and therefore invisible to this process. And if we are trying to understand, part of this is to get measures to evaluate people but the other is for learning.

And there are all kinds of things, all kinds of resources being used. The way we are currently measuring resources are invisible and, therefore, we can't learn what works and what doesn't. And that is an important limitation to this.

The other thing is, as we go back to increasing bundled payments, ACOs, capitated primary, patient-centered medical homes, we are expecting all kinds of interactions to take place among people in those organizations that are not billed interactions. Yet, we will have important consequences for how well they perform.

And if we are trying to understand

how well the health system does, we need to be able to understand whether those are taking place, whether they are taking place at different rates and what difference it is making in the care that people are receiving. So we need to think about how to do that.

The other thing is, as we move to bundled payments, if the past is any guide to the future, the billing and encountered data will become crappier again because we are not getting paid on it and we have got to worry about that in terms of the quality of the input for even the billing based measures that we are currently working with.

DR. NELSON: Thanks for those comments. Another gap, quickly, would be the indirect costs that the community and the employers are paying that we aren't capturing in our current approaches.

So Jennifer's card's up. Nancy, and then public comment.

DR. GARRETT: Just a quick

comment, which is in the provider world that I am in right now there is a real divide between quality measures, between kind of the quality type of work and type of data, and the financial data. And I think that is something that we are really going to have to evolve towards. And I think NQF could play a leadership role there. There just tends to be a real segmentation.

And I went back last night and I emailed to try and figure out where our Medicare spending per beneficiary reports are. And the quality person got them and sent them over to finance and they were sitting on their desk and no one had really looked at them. So that is what I will be doing tomorrow.

So maybe you need to change your name to National Value Forum instead of National Quality Forum.

DR. NELSON: Public comment, anyone in the room or on the phones. Not in the room. Do we have anyone on the phones?

MS. TIGHE: Operator? Operator, is there anyone on the phone who has a public comment?

OPERATOR: To make a comment at this time, please press *1. There are no comments.

DR. NELSON: Okay, so Ashlie and Lindsey, next steps. Wrap up.

MS. TIGHE: Okay, so next steps staff will go back and take the many recommendations and comments you made and try to synthesize this into something that is meaningful and understandable. And we will send that back to you in late June, early July for your review and your input. We will be posting it for a public and member comment period for 30 days, starting July 9th.

And then after we get all of those comments, which the predictions were hundreds, we will be convening you all via conference call to review the comments and make any updates to your recommendations.

After that, it goes through the rest of the process to NQF member voting, CSAC, and then ultimately Board of Director review and appeals.

MS. WILBON: So actually after all that, we actually finished on time. Can you guys believe that?

(Laughter.)

MS. WILBON: One minute early. Again, just a huge thank you to everyone for coming out. I know a lot of you came from very far away. And were really happy with having such an amazing group of people around the table. And so we are really excited about going forward and seeing how other people feel about our work today. They are as excited about it as we are.

So again, we will be following up via email. Either Lindsey or I will be in touch with further information. Again, if you can fill out you surveys, that would be great.

Thank you and safe travels.

(Whereupon, at 2:27 p.m., the foregoing proceeding was adjourned.)

B

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 Steering Committee

Before: NQF

Date: 05-09-13

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