NATIONAL QUALITY FORUM
+ + + + +
COST AND RESOURCE USE
STEERING COMMITTEE
+ + + + +
WEDNESDAY
MAY 8, 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:

HELEN BURSTIN, Senior Vice President, Performance Measures
ANN HAMMERSMITH, General Counsel

TAROON AMIN, 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

CRAIG CAPLAN, Centers for Medicare and Medicaid Services

JAYNE HART CHAMBERS, Federation of America Hospitals

GREG POPE, RTI International (by teleconference)

EUGENE RICH, MD, Mathematic Policy Research

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

JOHN SHAW, Next Wave

KIMBERLY SPALDING BUSH, Centers for Medicare & Medicaid Services

MARY WHEATLEY, Association of American Medical Colleges

SAJID ZAIDI, Acumen LLC

C-O-N-T-E-N-T-S

C-O-N-T-E-N-T-S (CONTINUED)

C-O-N-T-E-N-T-S (CONTINUED)

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

8:42 a.m.

CO-CHAIR NELSON: Good morning.

We would like to welcome you today for the Cost and Resource Use Committee deliberations, and we are very pleased that everyone is here.

As we think about how to operationalize the three part aim and the National Quality Strategy and the Affordable Care, measures of cost and resource use are clearly important. And so, we appreciate you taking the time today to be here with us.

CO-CHAIR PENSON: Thank you very much for coming. I look around the table and there are faces I know and some I don't know. So, it is good to see you people we have worked with before.

(Participant asks him to use the microphone.)

CO-CHAIR PENSON: I am sorry, I am talking into the microphone. I will talk

louder. I am from New York originally, so it shouldn't be a problem.

Anyway, having done this process before, I think all of you know, for those of you who have done it before, you know it can be a little bit challenging. But, that being said, I think these measures, we only have two. So, that will at least make the workload a little easier. I look forward to good discussions.

Taroon, I don't know if you want to add some comment.

MR. AMIN: Yes, I would. Thank you, David.

Again, thank you all. I really appreciate all the hard work. There is obviously a lot of materials that we sent out. We understand that that is a lot of work that you are doing on volunteer time. So, again, we appreciate all that work.

This is, obviously, not the end of the work. We will be going through, as

Lindsey and Ashlie will be talking through, you know, a comment period and going to our CSAC and to the Board for approval.

I just wanted to highlight a few components as we go forward here. Clearly, this is one of the nation's and, clearly, one of NQF's most high-profile projects this year. Because of that, obviously, we want to just stress a few components as we kind of move forward here.

The first, as you noted, we had a number of public comments that fed into this Steering Committee meeting. The role of public comments is obviously very important in the Committee discussions because, obviously, we are maintaining sort of a multi-stakeholder perspective here and all of the perspectives that feed into this discussion should weigh into the Committee's discussions.

So, again, making sure that we are weighing the inputs of the public, not only here in terms of the pre-member input, but

also evaluating those public comments as they arise during the Committee discussions. So, we will have various points during the Committee deliberations where we will ask for public comments and facilitate that discussion as we go.

The final Committee preliminary recommendations on this measures will also go out for public comments, in which we will ask you to evaluate and to think about those comments in the spirit of your deliberations.

The second is thinking about the fact that, obviously, these measures are, in the grand scheme of NQF measure portfolio, very complex in terms of the risk adjustment and their construction, and, obviously, issues of use are important here as well.

So, these conversations are highly technical. In a lot of ways, we have seated this Committee because of your technical expertise in various different areas. But it is also important to maintain the fact that

this organization sort of prides itself on the fact that we are multi-stakeholder, and the work of this organization will sort of feed into various different sectors of the healthcare landscape.

In particular, we should make sure that these discussions, once we have sort of digested the technical discussion, we can translate that into sort of a level of discussion that could be easily digestible by all stakeholders. And again, that is one piece.

And then, the second is to make sure that we look across the table during our introductions, recognize that we have various different perspectives represented here, and all those perspectives are valued at this table, and not to necessarily move the conversation into a technical domain, in which not everybody can participate.

So, if you do raise technical concerns on the measures, which would be valid

and appropriate for our discussions, please make sure that you kind of translate those discussions into a position that is understood and able to be translated to all broad stakeholders. So, I think that was the second piece.

And I think that was all I had to say. So, again, thank you all for being here, for all of your work that you have already put into this effort, and for the work that you are going to be putting in in the future.

CO-CHAIR PENSON: Thanks.

So, looking at the agenda, the next thing is to do introductions and, then, do disclosures. So, I would propose we go around the table, introduce ourselves, maybe say a little bit about where you are from, what stakeholder you may represent. And then, we will do the disclosures.

MS. HAMMERSMITH: No, we will do the disclosures at the same time.

CO-CHAIR PENSON: Oh, at the same

time. Okay, great. Okay. Thanks. Sorry.

MS. HAMMERSMITH: Hi, everyone.

I am Ann Hammersmith. I am NQF's General Counsel.

As was just said, we are going to combine introductions with disclosures. I see some familiar faces, so I know you have heard me say this about a thousand times, but I will give you a few reminders before we go around the table and you disclose.

You should have received a form from us where we asked you a number of questions about your professional activities, including your current position, any consulting, and so on. We culled through those as we seated people on the Committee. For members who have been seated, we like to do an oral disclosure in open session in the spirit of openness and transparency.

When you disclose, please tell us who you are, who you are with, and if you have anything to disclose. Please do not disclose

your entire CV because that will take a very long time, indeed, and we want you to be able to get to the work you are supposed to be doing today.

So, what you should be disclosing are things that you believe are relevant to the subject matter of what the Committee is going to talk about today. We are particularly interested in your disclosure of relevant consulting arrangements, research, or grants.

I also want to remind you that you serve as an individual. Sometimes people, going around the table, meaning well, say, "I am So-and-So and I represent the American Academy of" fill in the blank. And actually, you don't. You sit as individuals. You are here because you are experts. So, you do not represent the interest of your employee or of any organization that may have nominated you for service on the Committee.

And finally, I also want to remind

you that, because of the unique nature of what we do here, you may have something to disclose that is not a financial disclosure. Some people will say, "I don't have any financial conflicts," which is great, but it is possible that you could have something else that you should disclose where no money changed hands. For example, service on a committee where the subject matter is relevant to what you are going to be doing today.

So, with that, I would ask you to go around the table, tell us who you are, who you are with, and if you have anything to disclose. And we will start with the Chairs.

So, start with you.

CO-CHAIR NELSON: Gene Nelson. I am on the faculty at Dartmouth Medical School in the Dartmouth Institute, where some of the research involves Accountable Care Organizations and value measurement, and we have research and contract in these areas.

I am on the faculty of IHI and

have been working with them over time around the triple aim idea and its measurement.

And I am a founder of a company called Quality Data Management that is involved in quality measurement.

CO-CHAIR PENSON: Thanks.

My name is David Penson. I am a urologist from Vanderbilt University in Nashville, Tennessee. I am also a health services researcher. And obviously, I receive a salary from Vanderbilt.

I also have a consulting arrangement with the American Neurological Association as their Health Policy Chair.

I do research and have funding from various federal institutions, including NCI, AHRQ, and now PCORI, thank God. And I also have industry relations in terms of research agreements with a company called Dendrion and Medivation and Estelles.

MEMBER MARCINIAK: Good morning.

I am Martin Marciniak. I am with

GlaxoSmithKline. I am a senior leader at GlaxoSmithKline, and I have an equity position there as well.

I am the leader of the U.S. Health Outcomes Organization. It is an organization that has work that spans from quality of life to economic endpoints, both in clinical trials, observational data, and quality and economic measures.

MEMBER EAMES-HUFF: Good morning.

My name is Jennifer Eames-Huff. I am with the Pacific Business Group on Health. I am the Director of Advancing Policy there.

I also hold a role, being Director of the Consumer-Purchaser Disclosure Project.

MEMBER BECKER: Good morning.

I am Larry Becker. I work for Xerox Corporation in benefits and other things. You should know that Xerox Corporation also owns Buck Consulting.

I am on the Board of the National Quality Forum, and I also serve on the Board

of Governors of PCORI, the Patient-Centered Outcomes Research Institute.

MEMBER DAMBERG: I am Cheryl Damberg. I am a health services researcher from the RAND Corporation.

I have a number of contracts and grants that are in the value-based purchasing space. I just recently finished a project for the Office of the National Coordinator on identifying candidate measure concepts around efficiency and resource use.

I also sit on the eQuality Measures Work Group for the Office of the National Coordinator, and they have been considering inclusion of efficiency measures.

And I am doing evaluation research around the impact of value-based purchasing, and some of those programs include efficiency measures embedded in them.

MEMBER NAESSENS: Good morning.

I am Jim Naessens. I am health services researcher at Mayo Clinic. I am also

involved with the Value Analysis Program within the Clinic.

Also, I have a subcontract research grant with Minnesota Hospital Association, an AHRQ grant, and also serve on some Minnesota Department of Health advisory committees in terms of value and cost measurement.

MEMBER CLARK: Hi. I am Mary Ann Clark. I am Senior Vice President with a brand-new company called Interalign. We help work with hospitals to improve costs and quality. I am doing a lot of work right now with looking at various cost and quality measures for hospitals.

MEMBER WONG: Hi. I am Herb Wong. I a Senior Economist with the Agency for Health Care Research and Quality. The agency is a federal agency that conducts and supports research on health services research. And I have nothing to disclose.

MEMBER STEPHANSKY: Hi. I am Joe

Stephansky. I am with the Michigan Health and Hospital Organization, a trade association as well as a lobbying advocacy organization.

I have worked very hard to maintain plausible deniability in everything that I do. So, I have no association with any grants or other organizations.

MEMBER REDFEARN: Good morning.

I am David Redfearn. I work for WellPoint. The work I have done for WellPoint over the years has been focused on physician cost-efficiency profiling, risk adjustment, episode-of-care methodology. I have no conflicts to report.

MEMBER GELZER: Good morning.

I am Andrea Gelzer. I am Corporate Chief Medical Officer for the AmeriHealth Caritas family of companies. We do Medicaid-managed care and are also in the dual-eligible space.

I am on the AHRQ National Advisory Council, and I am also on clinical leadership

committees at both AHIP and Medicaid Health Plans of America.

MR. WOLFSON: I am Danny Wolfson, Executive Vice President and Chief Operating Officer of the ABIM Foundation. We are an independent corporation from the American Board of Internal Medicine, but we have interlocking boards. I run the Choosing Wisely Campaign. We have a grant from the Robert Wood Johnson Foundation to spread the campaign across the country.

MEMBER LATTS: Good morning.

I am Lisa Latts. I am a physician currently working as an independent consultant in the quality arena, formerly with WellPoint, where I would turn to David where I needed anything to do with smart physician efficiency.

My current clients include organizations such as the Colorado Health Exchange, where I am working on their quality strategy; pharmaceutical and biotech clients,

working on quality and payer integration, and the University of California.

MEMBER TSANG: Good morning.

Tom Tsang. I am Executive Director at Merck Pharmaceuticals. I work in the Office of the Chief Medical Informatics and Innovation Office, and I work on health IT partnerships and strategy for the company. I was formerly Medical Director for Meaningful Use at ONC.

MEMBER NEEDLEMAN: I am Jack Needleman. I am a professor of health policy and management at the UCLA Fielding School of Public Health. I am also the Associate Director of the UCLA Patient Safety Institute.

I occasionally sit on the Quality Committee for the American Academy of Nursing, and currently I am working with the UCLA Center for Health Policy Research and UC-Berkeley Labor Center on studying the implementation of the Affordable Care Act in California. And no other disclosures.

MEMBER ASPLIN: Good morning.

My name is Brent Asplin. I am Chief Clinical Officer and President of the Fairview Medical Group, part of the Fairview Health Services, in Minneapolis. I am an emergency physician, Past Chair of the American College of Emergency Physicians' Quality and Performance Committee. Fairview is one of the pioneer AC organizations in the U.S. And I have nothing to disclose.

MEMBER YANAGIHARA: Good morning.

I am Dolores Yanagihara with the Integrated Healthcare Association in California, where I lead the performance measurement and analytics area. We use efficiency measures.

And the only thing I have to disclose is that I am on NCQA's Efficiency Measurement Advisory Panel.

MEMBER GARRETT: Good morning.

I am Nancy Garrett from Hennepin County Medical Center, which is a safety-net

care system in Minneapolis. I lead analytics there.

And I have a background doing provider profiling at several health plans as well as working with Minnesota Community Measurement through the Board and some of our efforts in the community to measure both cost and quality.

And I don't have any disclosures.

MEMBER GIFFORD: Hi. My name is David Gifford. I am a geriatrician. I am the Senior VP for Quality and Regulatory Affairs at American Healthcare Association, which represents nursing homes and assisted living. I am also on the faculty at the School of Public Health at Brown.

My wife works for the Medicaid Program. She is the Medical Director in Rhode Island.

And I have a 401(k) that God knows what they invest in, probably things I don't support, and probably things I don't support

here, but something out there.

(Laughter.)

Other than that, no disclosures.

MEMBER McHUGH: I am Matthew McHugh. I am faculty at the University of Pennsylvania School of Nursing, Center for Health Outcomes and Policy Research. I am a Senior Fellow at the Leonard Davis Institute for Health Economics.

I am a health services researcher. So, predicatively, I have funding from various federal and nonprofit agencies like the National Institute on Aging, the Robert Wood Johnson Foundation, the National Institute for Nursing Research. And some of that is associated with looking at value-based-purchasing-oriented outcomes.

MEMBER PARE: Hi. I am Carolyn Pare. I am President and CEO of the Minnesota Health Action Group. We used to be the Buyers Healthcare Action Group based in Minneapolis, Minnesota. It seems to be Minnesota is pretty

well-represented here.

I also do work, I sit on the Consensus Standards Approval Committee of the NQF. I sit on the Standards Committee for NCQA, and I am on the California Health Benefits Review Program as well. I have no conflicts.

MEMBER WEINTRAUB: Good morning.

I am Bill Weintraub. I am Chair of Cardiology at Christiana Care in Delaware and professor of medicine at Thomas Jefferson University. I am a cardiovascular epidemiologist and outcomes researcher.

I have grant funding from CMS and the National Institutes of Health, several different agencies.

I am the President of the Great Rivers Affiliate of the American Heart Association, incoming Chair of the Data Standards Task Force of the American College of Cardiology and American Heart Association. I am also on multiple American College of

Cardiology committees, including Informatics, Data Standards, and NCDR Board. I do consulting for Pfizer and Daiichi Sankyo.

MEMBER WALKER: Good morning.

I am Lina Walker. I direct the health policy team at the Public Policy Institute in AARP, and I lead the team's work on Medicare and Medicaid and coverage issues for the under-65 population. I have nothing to disclose.

MEMBER RYAN: Hi. Andy Ryan. I am assistant professor at Weill Cornell Medical College.

I have grants from AHRQ and RWJ to evaluate value-based purchasing and quality profiling in Medicare, and I have a consulting relationship with Ecometrica to evaluate Medicare's bundled payment demo.

Thanks.

MS. HAMMERSMITH: Thank you.

Are there any Committee members on the phone? Specifically, Stanley Hichberg?

(No response.)

No? Okay.

Thank you for those disclosures. Do you have any questions of each other or of me, based on the disclosures this morning, anything you would like to discuss?

(No response.)

Okay. Thank you. Have a good meeting.

DR. BURSTIN: I would add my welcome as well.

I am Helen Burstin. I am the Senior Vice President for Performance Measurement at NQF.

Thank you to all of you for coming. I know this was a pretty significant lift, lots of materials, and I am looking forward to a great discussion.

Thank you.

MS. TIGHE: Hi. I am Lindsey Tighe. I am the Project Manager who has been emailing you consistently.

(Laughter.)

I figured I would, first, just run through the papers that we printed out for you today and just let you know what you collected out there.

So, we have the agenda. We have got the rosters and bios for everybody on the Committee. We have got a handout of the Resource Use Measure Evaluation Criteria. We have a set of slides for each measure that the measure developers provided.

We will run out and grab them for you. They are on the table out there. Yes, we can bring some in and get them passed around to everyone.

They are reference materials. Most of them you already have except for the slides and memos that the measure developers have provided, which, when we actually are discussing their measures, they will walk through those with you. So, we have got one set of those from each measure developer.

We have got a table that lays out the HealthPartners measure that we sent out to you, the total resource use PMPM measure with the total cost measure from CMS. We will use that later this afternoon when we have the harmonization discussion of related measures.

And then, we have got the compiled preliminary evaluations for you all, so you are able to reference that during our discussions today.

So, hopefully, those are helpful. Most of them you already have, but we just wanted to have handy references, as we have our discussion.

MS. WILBON: Just for the sake of introduction, my name is Ashlie Wilbon. I think I was able to speak to everyone this morning. Some of you I know; most of you I know.

I just want to thank you guys again for coming. It is great to have such an amazing group of people around the table to

discuss these important measures.

Evan Williamson is our Project Analyst. He will be walking around the room helping us with various things today. So, he is an important team member I wanted to recognize.

And again, we are happy to have you guys here.

MS. TIGHE: Yes, my hesitation was Evan is actually running the slides.

Okay. So, I will just briefly run through a little bit of the meeting objectives and the project background. And then, Ashlie will give you all a refresher on the measure evaluation criteria.

So, our objectives for today are review both of the measures that were submitted against the measure evaluation criteria and, then, make a recommendation for endorsement.

This afternoon we will be discussing the two related measures, the

HealthPartners one and the CMS one, to determine whether or not they need to be harmonized and, if so, what elements should be harmonized.

And the third objective is broken into several pieces. This afternoon we will also just be asking you all to give some broad input to the Measures Application Partnership on recommendations for the measures. For those of you who are less familiar with the MAP, it is a group that we have NQF that makes recommendations for application of the measures into federal programs. And so, we will just be asking you all after you evaluate these measures to just give them some broad considerations, as they consider these measures for federal programs.

Tomorrow we are going to have a broad discussion of the implications for specifying and endorsing measures with multiple risk adjusters. We have a measure that is currently endorsed where we will be

looking at the possibility of adding on additional risk-adjustment models. And we are trying to really establish NQF policy going forward on how we handle that.

We will also be asking for your input on inclusion of attribution guidelines within the cost and resource use measure submissions. As you all saw, it is currently an appendix to the measures. It is not something that we evaluate in our criteria. And so, we will have a conversation about how we should use that information going forward as we evaluate cost and resource use measures.

And then, we will also have a broader discussion about moving toward efficiency and value measures, where you take these cost and resource use measures and you pair them with quality measures, so that you can really start getting at the value of healthcare services.

Okay. Moving into the project background, as you all are well aware, there

are two measure submissions to this project. As I mentioned earlier, we will be looking to explore the potential for harmonization and discussing the application and use of the risk-adjustment models within the cost and resource use measures. This is all under the Phase 1 of the Cost and Resource Use Project.

We have a planned Phase 2, which will be condition-specific, episode-based measures. As it says, detailed "TBD". We are expecting that sometime this fall.

So, we have an eight-step consensus development process for measure endorsement. This is currently the standards review step of that, Step 3, where we have Committee review of the measures and recommendations for endorsement.

After this, staff will compile a draft report which we will put out for public and member comment. So, we will be getting multi-stakeholder input on the Committee's recommendations for endorsement. That is a

30-day period.

After that, we will bring those comments back to you all for consideration to determine whether or not you agree with your preliminary endorsement recommendations and want to continue to recommend the measures that you have chosen or whether some comment has come up that causes you to reconsider the endorsement recommendation.

From there, we move into NQF membership voting, a 15-day period where only the NQF membership can vote. Once we get the results of that, this moves on to the Consensus Standards Approval Committee. So, Carolyn will be shepherding these measures forward from today on.

They will give a final review of these measures at their meeting. And then, from there, it will go on to the Board of Directors for final endorsement, which Larry will continue on at that point.

Once the Board makes their final

decision, we do have a 30-day appeals period where anyone can submit an appeal if they feel that they are materially affected by the endorsement of the measure.

This is our timeline. So, today the Steering Committee in-person meeting, May 8th and 9th.

Over the next two months, staff will be working hard to capture everything that we discuss today, compile it into a draft report, share that with you all to make sure that it is okay, and put it out for the NQF member and public comment period in early July.

So, this project runs through the fall. It ends in late October and goes through all those steps that I mentioned before with commenting, vote, CSAC, and Board of Directors approval and appeals.

I will stop there. If there are any questions on the background, jump in.

(No response.)

Otherwise, Ashlie will go over the measure evaluation criteria.

MS. WILBON: So, we have been through this, the criteria, a couple of times. You guys have already done your preliminary evaluation. So, I am not going to belabor this, but just really try to highlight some points that we think are important to consider again, as you guys are now kind of convened together as a group.

So, again, the four major criteria around importance, scientific acceptability, feasibility, and usability. Just to note, as we go through the criteria today and the discussion, both importance and science acceptability are must-pass criteria. If we get to the point of importance and it does not pass importance, we wouldn't go through the remainder of the criteria, and the same with scientific acceptability. So, those are kind of threshold criteria, if you will, and we will go through them in the order that you see

listed.

So, again, the subcriteria that you evaluate in your preliminary evaluations are really just to help think through how the measure would demonstrate meeting the overall criteria. And the criteria really developed around best practices in measure development.

So, when you develop a measure, first, you want to start with the concept. Is the concept important to measure? And then, work your way down through feasibility.

And so, obviously, the purpose of having experts here is that there is no black and white with these. These are all kind of a matter of degree in which we use experts to kind of help us make those decisions on whether or not the measure meets the criteria. So, that is what you are all here for. So, we need your expertise and judgment to determine, you know, the degree of demonstration of the criteria.

So, again, the first criteria of

importance to measure and report, there are three subcriteria. We will walk through whether or not the developer has demonstrated that the measure focus is a high-impact area to measure.

We will walk through whether or not they have demonstrated information on there is an opportunity for improvement, whether or not there are variations in disparities in care delivery in this particular measure topic. And then, we will talk through whether or not the intent of the measure has been sufficiently supported with the resource use categories that are specified in the measure.

Right, Taroon is reminding me in my ear that we do have a Criteria 2c, which we will get to -- actually, you will see here on this slide -- for stratification for disparities. It is kind of our Lone Ranger out here in the reliability and validity under scientific acceptability.

What we will be doing is kind of including that discussion of stratification and disparities under 1b and the importance criteria, which discusses variation in care or disparities in healthcare delivery. So, for those of you, I think we communicated with everyone via email who had been assigned to C on one of the measure discussions that we will be kind of asking you to discuss any issues around disparities and the importance discussion.

So, the second criteria is scientific acceptability of the measure properties, which focused on reliability and validity. Within each of these areas, again, both are must-pass criteria. We will be discussing the specifications and the testing related to the measure specifications and whether or not they have demonstrated reliability and validity.

Reliability is looking at the preciseness of the specifications and

reliability testing, either at the data element or the measure score level.

The validity subcriteria is focused, again, on the specifications and whether or not they are consistent with the measure intent. We will be looking at the actual validity testing, at the data element or the measure score level. It also includes the testing and justification of exclusions, the risk-adjustment model, the identification of differences in performance, and comparability of data sources and methods, which I think we discussed on one of our tutorial calls; that because there is only one data source specified, that that will probably be a moot point, but it is part of the criteria.

So, just to point out again, I think you have seen our little house before. Just to kind of reiterate how we have broken down the specifications of the resource use measures into these different buckets of

specifications or modules, if you will, general methods and kind of data protocol and how you prepare the data. And then, looking at clinical logic, how the specific patients are categorized by different clinical characteristics.

You also have your construction logic, which is more kind of setting up the measure, how the episode or the timeframe is defined, and how claims are assigned to the episode.

The resource use service categories are how the claims and codes are categorized. So, for example, pharmacy claims or E&M visits or DME or other types of categories that you would use.

And then, the adjustments for comparability, which is going to include the costing methodology, the risk-adjustment approach, the stratification, and then, finally, the reporting piece, which would include different guidelines for the measure

that you would use when you would go to report, including peer grouping. Right now, we also have attribution there, benchmarking, and other things like that.

So, reliability and validity, again, are not all-or-none properties. They are matter of degree, which we have asked you to rate on a scale of high, moderate, to low or insufficient. And obviously, they can vary with different conditions of the measure.

And again, just kind of basic principles. In order to be valid, the measure must be reliable, but, conversely, reliability does not guarantee validity.

And I will just note that we will be asking you to vote on overall reliability and overall validity. We don't ask you to vote on overall scientific acceptability. If you vote that the measure is either high or moderate on reliability or validity, the measure will pass scientific acceptability automatically. If you vote low on either one

of those, the measure would not pass scientific acceptability.

So, again, reliability is repeatability or precision of the measure, and validity is the correctness, the accuracy.

So, some threats to reliability, these are some things you will be thinking through potentially. Hopefully, you thought through these in your preliminary evaluations, but they are obviously on the table again today.

Ambiguous measure specifications. So, not really understanding how codes are being assigned or divided up into categories. Small case volume or sample size or rare events, those can obviously affect the precision of the measure score. And then, random errors, either for coding or for missing data.

Some threats to validity. Conceptual, so if the measure focus is not relevant or not strongly linked to some

particular cost that the measure is intended to measure. On reliability, patients that are inappropriately excluded; differences in patient mix, and measure scores generated with multiple data sources or methods. Again, we won't really encounter that with these two measures. And then, systematic missing or incorrect data.

So, measure testing is a huge part of how the measure developers are to demonstrate reliability and validity. We are looking for them to demonstrate the reliability and validity of the measure as specified.

So, another note for today is that we are looking and evaluating the measure in front of us. I know a lot of times we like to think about how we could make the measure better or if they just did this or if they just added this thing or tweaked this, and we are really going to focus on the measure as it is specified in front of us today.

So, this is just some additional guidance on the testing that we provide to developers: empirical evidence, basically, we are expecting them to do testing. And again, we are looking at the reliability and validity testing on the measure as specified. That would include the preciseness of the specifications and whether or not the specifications are consistent with the measure content.

We are not prescriptive to developers on which type of testing they need to use to demonstrate reliability and validity. We leave that to them. Obviously, they have developed the measure and are going to know the best way to be able to demonstrate that based on the specifics of their measures. So, what you see in front of you is what the developer has decided to put forth in terms of demonstrating those two criteria.

Specifically, I think we had maybe a couple of conversations on maybe the

tutorial calls about this. When we start looking at R-squared values or correlation values, we don't have necessarily set thresholds to say, okay, an R-squared of -- I don't know -- .7, or whatever, is high or moderate. That is really what we are having you here to help us determine. A lot of times those values are in the context of the measure construct and how it is being used, and so forth.

So, we don't set specific thresholds for that. So, within your expertise and your realm of understanding how statistical values work, we are hoping that you will bring that and help us make a decision on whether or not those are valid. And again, the testing does not replace the need for expertise and judgment, again, which you are here for.

I won't spend a lot of time on this, but these are some of the strategies that we have implemented in our process to

mitigate the burden of testing for developers. So, rather than having to submit testing at both the data element and the measure score level, we allow them to do an "either/or". I will note that the high rating on reliability and validity testing is contingent on them doing both data element and measure score testing. So, if they have done one or the other, they would only be able to score at the highest of moderate of those.

Again, we allow them to do testing on samples. They can use prior evidence, which doesn't necessarily apply here, but empirical evidence of data element validity, and -- I'm sorry -- a separate reliability data element is not required. So, if they demonstrate data element validity, they don't necessarily need to do reliability testing because inherently the data elements would be reliable if they are valid.

And face validity is accepted if it is systematically assessed, which is not

really applicable here.

Reliability assessing, these are just a couple of examples of how people could demonstrate or how developers would be able to demonstrate reliability testing at the data element level, inter-rater reliability, or internal consistency and reliability for multi-item instruments. Again, that will be dependent on the specifications and the data used to construct the measure.

Next. Go back one more.

At the measure score level, what we are looking for is for them to demonstrate that the measure is reliable and the proportion, the variation of the scores due to systematic differences across the measure entities in relation to random variation. So, again, this would be kind of a signal-to-noise analysis at the measure score level.

For validity testing, at the measure score or the data element level, we are looking for them to demonstrate that the

measure score and data elements are correct and they accurately reflect the cost-of-care resources provided. An example for how they might be able to demonstrate this for a resource use measure would be correlating with measures that are deemed to be related in some way or testing the differences in scores between those groups that are known to either differ or be similar to, have similar utilization or cost-of-resource use with the measure that is under review.

So, we will be going through this. As we get to the measure evaluation, we will be having some kind of discussion points or questions, things that should be considered as you are evaluating the different criteria.

Specifically for testing, obviously, we will be looking at appropriateness, scope, and acceptable norm. So, did they use the appropriate type of testing to demonstrate their measure was reliable and valid based on the way the

measure is constructed? Is the scope adequate? So, did they include an adequate sample? Did they consider an adequate number of sample entities, the number of patients, and the representativeness of the sample?

And then, finally, whether the results were in acceptable norms; again, back to that issue of kind of the values of the statistical tests that are performed, we will be looking for you to make that determination.

So, risk adjustment. We are going to have quite the discussion on this tomorrow. I am not going to spend a whole lot of time on this. But, essentially, it is used as a case-mix adjustment in the measure, controlling for patient factors that influence outcomes in the measure.

And generally, what we are looking for here is that the factors that are included in the measure are present; the patient-level factors that are included in the model were present with the patient before the measure

began. So, we are not looking for them to include patient-level factors that impacted the patient during the episode of care. So, I don't want to get into a whole lot of detail about how that could get circular where the severity of the patient is included based on things that happened during the episode. So, that is basically what that is getting at.

A couple of other things about risk adjustment. We general, as a rule here at NQF, don't look for them to include things that relate to disparity. So, race and ethnicity should not be included in the model. We would rather see those things be stratified, so that we can actually identify where those differences are and focus on where the improvements or additional attention needs to be focused.

And also, we don't expect that they would include structural characteristics for organizations. So, experience, whether it is a teaching hospital or equipment

availability, or things like that, in the risk model.

So, I have been through some of this throughout the presentation, and I won't focus on this. We will be guiding you guys through this. As we get through the evaluation, I think it is a lot easier to apply it rather than me just talking at you.

But, again, in order to get a high for reliability and validity, we are saying that the specifications are precise and accurate and that they have done testing that you feel is adequate and appropriate at both the data element and the measure score level.

And a moderate score would be, again, that the specifications are in line with what you think they should be, they are precise and accurate, and that they have done testing at either the data element or the measure score level.

And again, low would be that there are ambiguous specifications and that the

testing or the scope of testing has been inadequate or inappropriate at some level.

And we won't go through this. This, again, is the algorithm we will use to determine whether or not the measure has met scientific acceptability.

Feasibility is the third criteria. We will be looking how feasible it would be for an implementer to actually take the measure and run it or use it at their particular system or facility; whether or not the data are generated during care processes; whether or not they are available electronically, and whether or not a data collection strategy can be implemented. I think this is probably one of our more straightforward ones.

Usability and use, which also we anticipate will be an interesting discussion, focuses around four subcriteria, the first being on accountability. We are looking to see whether or not the measure is actually in

use at this point in time for, one, accountability application. If it is not, we are asking for them to provide a plan or some indication of how they anticipate the measure will be used going forward.

In 4C, we are looking to determine whether or not that the benefits of the measure would outweigh the harm. So, for example, if they had identified some unanticipated, unintended consequences in either the testing of the measure or through use of the measure, if the measure is actually in use, that we would ask you to look at those unintended consequences and determine whether or not the benefits of using the measure would outweigh the harms that would be imposed with those unintended consequences.

And then, finally, with 4d, we are asking you to determine, based on the specifications that you reviewed, would the results being shared with other stakeholders be transparent enough for people to kind of

deconstruct that and understand what is being measured and what the results actually mean?

And finally, the last part of the discussion that we will have kind of, hopefully, at the end of today will be on comparing related measures. We have a measure that we endorsed in our last resource use project from HealthPartners that we have identified as a similar measure to the total-cost-per-beneficiary measure that you will be reviewing today.

And once you go through the measures, the two measures that are in front of you today, make your recommendations, we will kind of reconvene and decide, have a discussion on the comparison of the previously-endorsed measure with the measure that is on the table and see where there may need to be alignment. And we have the developers here from both sides who will help us walk through that discussion.

Again, the rating scales for

today: high, moderate, low, and insufficient. And we will be kind of helping you walk through that today as we go.

Lindsey is going to kind of talk us through some of the broad concepts we are going to be asking you guys to think through today as you make your recommendations.

MS. TIGHE: Sure. So, this graphic just really demonstrates kind of the related relationship that all of our involved parties have here: that we really want this to be a two-way street where the measure developers and the measure users are informing you all, and you all are informing them with your recommendations. And the same goes for the Measures Application Partnership Work Groups and Committees.

So, we are really going to be looking to emphasize these relationships and just really provide that opportunity for feedback to be provided throughout all the groups, so that there is kind of consistency

in the work that we are doing.

So, that said, we just really want to highlight of the Steering Committee versus the MAP groups. So, the Steering Committee is evaluating the measures for broad use for performance improvement and accountability purposes. You are looking at whether it is useful or whether the specifications are appropriate for it to be used for performance improvement and accountability purposes.

The MAP is actually looking at whether it is appropriate for specific application and accountability programs. So, they are going to be looking at whether or not it should be used for reporting in PQRS or in value-based purchasing programs, something to that effect.

Did you have something to say, Ashlie? Okay.

So, what we are looking for here is we had a MAP meeting in December. They provided some input to the Steering Committee

for your consideration, as you are looking at the resource use measures.

Some of this is a little bit more broad and is related to the work that they are doing. So, it may be areas where you can provide guidance as you evaluate the measures today.

But they are looking for resource use measures to be linked with outcome measures. So, they are really looking to move toward that value that we discussed earlier, looking at the value of healthcare services provided.

And there will be a future MAP affordability family of measures that will be recommending specific quality measures to link with the resource use measures coming up.

They also recommended that the resource use approaches should align across populations and settings, using the same measure when feasible. So, again, just considering how the risk adjustment and

attribution methodologies could align across populations and settings, and assuring that the measures that we are recommending today do address the broadest target population possible.

And their final recommendation was that the resource use measures should be patient-centered. A way to do that would be that condition-specific resource use approaches should address multiple chronic conditions.

We also had -- oh, go ahead.

MR. AMIN: Lindsey, before you go on, I just wanted to just make sure that there are no questions about that. I know we are having some technical difficulties on the main slide deck here, which we are getting resolved.

But this is a very important issue around the difference between the MAP function and the Steering Committee function. So, Lindsey kind of articulated exactly what those

differences are. I mean, clearly, there is some informing between the two groups, but we want to make sure that this group understands its function, which is to look at whether these measures are appropriate for broad application and not specifically make recommendations during its evaluation around particular application of these measures in particular accountability programs.

Clearly, some of the measures in front of you are already in statute for particular programs. So, it is not like we are trying to say -- you know, clearly, you are going to have an opinion about that. There is going to be a particular time to provide that input, which will be after your evaluation of these measures.

So, we just wanted to make that clear, to the extent that it can be clear. If there is any discussion or questions about that, please, this would be the time to raise those concerns, questions, comments.

And I will hand it back to the Chairs on that.

CO-CHAIR PENSON: Great. So, Jack, you have a question.

What I propose, just going forward, is if you have a question, just put your tent up. So, that way, you don't have to raise your hand.

Jack?

MEMBER NEEDLEMAN: Okay. Yes, I am now confused. The resource use -- in the first iteration of this -- I was on that Steering Committee as well -- we had a long discussion about the difference between quality, between efficiency, which in fact was the linkage between quality and resources used, and the resource use measures. And we said the state of the art didn't feel like it was ready for efficiency measures yet, that one of the stepping stones for that was the development of effective measures of resource use.

We are being asked to endorse this set of measures on resource use. The MAP criteria just threw up basically said: no resource use measures until we have quality measures to go with them, so we can measure efficiency.

So, what are we doing? What does that criterion do to our reflection? What does that MAP concern do to our reflection on the appropriateness of endorsing this measure now?

MR. AMIN: Okay. So, actually, I think it is the slide after this that Lindsey was walking through the input. And many of you were part of those discussions. I don't want to call anybody out, but you are sitting on the MAP groups.

But the MAP made some recommendations to this Committee, meaning that, as you are approaching this exercise, think about the following considerations, and these are broad considerations: and those

considerations were around how can we get to -- it is really more of a forward-looking conversation. It goes back to sort of our evening discussion today and our discussions tomorrow, which is the fact that this group and its previous iteration of this group -- some of you were on that -- made recommendations about, first, let's get the resource use measures out there. And then, let's really come up with an approach to link the cost and quality domains to be able to get a real signal of efficiency.

And so, what the MAP group is providing back to you to consider is, as we are moving forward through the endorsement, post-endorsement of this, how do we move and actually get these two concepts to link? How do we actually make this happen? How do we move the measurement field to really understand how to link cost and quality in a way to really understand efficiency?

So, separate and apart, you have a

task in front of you. We want to separate -- there are two different tasks that are going to be in front of us during the two days that we are here.

There is the actual looking at these measures against the criteria that currently exists. We are not going to make changes to criteria in real-time here. That is not the purpose.

At the end of the evaluation of the measures, there will be some forward-looking discussions. Those include this issue around how do we actually get to a measurement space that actually does do the linkage between cost and quality, and what does that look like? So, that we can actually start moving toward that future. That will be part of our gaps discussion later on this afternoon.

And then, also, the forward-looking will address issues around, as we talked about this issue around risk adjustment

and attribution, which still need seem to be a major issue in terms in the field about how we conceptualize this going forward.

So, very specifically, the answer to your question, Jack, is that these are considerations; they are not criteria. And they are meant to inform your discussions, just as the public comments are meant to inform your discussions.

MS. WILBON: I will just add that, in the context of this discussion, they were looking to make recommendations of linking, when they make recommendations for specific programs, to say we would like this resource use measure used with X, Y, Z quality measures, and that they are looking to make those linkages as a part of their recommendations. That still requires the need for them to have an endorsed resource use measure to include in that recommendation.

So, the work of this Committee is still to focus on having reliable, valid

resource use measures that can be a part of that discussion. The MAP -- and Dolores and Gene can chime in on this -- really tends to give deference to endorsed measures. So, they would rather make recommendations for endorsed measures that are available that meet the needs of that program before kind of going outside the box and looking for what else is out there.

So, the work of this Committee in going through and deciding whether or not this meets the criteria is actually really important. It allows them to have something on the table to include in their recommendations for that. So, that is kind of, I guess, some more of the context that was in there.

And I can open it up to anybody else, Dolores or Gene, who have been sitting on MAP and were actually there for those discussions, to add to that.

CO-CHAIR NELSON: Yes, just very

briefly, this could be considered or these could be considered building blocks towards measures of efficiency or towards measurements of value, a building block in relationship to other building blocks.

CO-CHAIR PENSON: Can I ask a question related?

Oh, I'm sorry, go ahead, Joe.

MEMBER STEPHANSKY: I think some of the attempt to differentiate between what we are going to do here and what MAP does is my fault because of some questions that I raised earlier.

It really comes down to, since I am from the Hospital Association, one of the viewpoints that I want to make sure people understand is that hospitals are not going to look at that, this use of the CMS measure to hold hospitals accountable for things that happen 30 days after discharge. It may not be fair, and we don't want to really discuss that here. We want to discuss the measure and

leave the use of the measure by CMS to MAP. I think that is where we are trying to go.

MS. WILBON: And I would just say, Joe, I can take the blame off of you. We had actually anticipated that that would be, from other experience in other projects, that that was actually an important differentiation we needed to make. So, you are off the hook, and we recognize that is an issue. But there will be an opportunity to discuss that.

CO-CHAIR PENSON: Other questions?

(No response.)

So, I have a question related to the MAP and to our role specifically -- and I think it was mentioned in the next slide -- about harmonization. So, does that fall under the Steering Committee? Does it fall under the MAP? What is expected from us as far as harmonization goes, since I see there is a discussion about it tomorrow?

MS. WILBON: Harmonization is actually under the purview of this Committee

because you guys have the opportunity to really dive into the specifications. The MAP's discussions are at a much higher level. So, when we get into the harmonization discussion, we will be talking a little bit more about conceptual harmonization versus technical harmonization, which is really you guys are really under the hood, looking at the codes, how things are defined. That is really under the purview of this Committee.

So, we will be specifically asking you guys to make recommendations on harmonization, and we will do our best to kind of carry those forward to the MAP Committee to make sure they understand kind of the decisions that were made here. But that would be you guys.

MEMBER WALKER: Just a little more clarification on the role of the Steering Committee in evaluating this measure. Now you said that we should evaluate it for broad applications. But when you read the narrative

from the developers, they write it, target it for a specific application. So, are we in our evaluation supposed to distance ourselves from that narrative in which they describe how it would be useful for that particular specification, because we have to think about it or evaluate it more broadly?

MS. WILBON: That is a very good question. It gets a little tricky. Our criteria actually is for broad application. And so, it does get a little tricky when the submission is very specific to that program, and we are trying to kind of walk that fine line between us and MAP, yes.

And I am going to let Taroon respond further.

MR. AMIN: So, the criteria is written to really think about the measure in terms of broad application. There was a group that looked at the use and usability of measures in our portfolio, with the goal, with the lens of thinking that we don't want to

have a large number of measures in our portfolio that are not being used in the field, because that sends a signal, even if we have looked at the use and usability of the measure, and they are not being taken up, that that is a concern. You know, that is a concern; that should be a concern for us broadly.

And so, what we have asked the developers to do is to provide a plan for how these measures are going to be used and their specific plan for how these measures are going to be used, as a way to inform our thinking about the fact that these measures will be taken up.

The developers obviously have an intent of how these measures will be used. And so, we should think about that in the context of can they be usable and can they be taken up by a program, but not necessarily whether they are appropriate for that specific program.

So, that is where the difference is, right? I mean, it is still a little bit related. The reason why we asked for the specific program is that we wanted to make sure that they could actually be taken up and used in various accountability applications or quality improvement applications. So, that is the context in which you should evaluate that.

We will go through that in a little more detail when we get to the usability criteria. But is at least that clear enough for now? And if it is not, let's talk about that because, again, it is a really important question.

MS. WILBON: We are going to work through it together. Don't worry. It will be okay.

(Laughter.)

It is going to be hairy. We are recognizing that, and it is difficult, but we are going to work through it. So, bear with us. We will get through it.

MS. TIGHE: Next slide.

So, on April 25th, we had an all-member call with the NQF membership. The purpose of this was to really start getting our membership engaged upfront and have it feed into the Steering Committee review rather than it simply being a response to the Committee recommendation.

This took place during the time when we had a commenting period open on the two measures for the public and membership to provide input. So, again, we are really just looking to get there, that you all can consider it as we make our initial recommendations for endorsement.

And that input from them was along the lines of it was also very broad and it touched on some of what we just discussed as relates to the Measures Application Partnership, but just really looking at if the measures allow for identification of high-value target areas for providers to focus

efforts on for reducing costs. So, kind of getting at if the measures can be deconstructed, so that providers know where to focus.

Looking at how access to care can be assessed within the context of cost and resource use measures. This one, again, may be a little bit more of a consideration for measures as they are developed. How are vulnerable populations addressed within the context of the measures being reviewed? What is needed to move toward assessing value of services? So, this will be a conversation that we have later today in the measure gaps discussion.

And then, how can drivers of cost be tied to quality of care to allow for identification of an ideal state? So, again, part of that discussion later today.

MS. WILBON: So, this slide, again, gets to Gene's comment earlier about building block. To kind of center us where we

are focused today, again, it is in that resource use box. But we completely recognize that this is a building block towards getting to value and efficiency. This is just one piece of the puzzle, and we just need to make sure this piece of the puzzle is valid, reliability, and that there is agreement that this is a valid indicator of cost and resource use. So, that is really what we are getting at today.

There was work done, I guess back in 2009, on defining and episode of care and applying that to how cost and resources should be measured. We like to call this the caterpillar model, to show how an episode of care for a particular patient, if you are looking at trying to develop measures in a patient-centered manner, how measures might be applied to that episode and to be patient-centered and apply across the episode.

So, again, you start with a population. You move into the evaluation and

management phase. And then, you have a followup phase.

And really, just kind of be thinking about this episode of care in your mind as we are evaluating measures and how these measures can be applied across the episode of care and their patient-centeredness.

And I won't spend time on this. We went through this in the orientation and tutorial calls. But some of the issues that we know that have come up before when evaluating these measures, and we anticipate that some of them will come up again. Just to kind of have them in your mind, and we have been here before somewhat. We expect some of these issues to come up again.

Again, the linking quality and cost issues; the costing approach and how that is defined and what implications that has for measurement. We will be talking a lot about risk adjustment. The impact of carveouts on

the outcome of the measure; not so much reliability and validity at the individual physician level, but, again, this whole public and private sector alignment we will be getting to a lot in the harmonization discussion.

So, again, just some things for your broad consideration. There are a lot of moving pieces to this going on in the landscape right now. We just kind of what to make sure that everyone is kind of aware of what is going on and the background.

So, we are getting ready, actually, to move into actually evaluating the measures, which you have all been waiting for. So, hopefully, you guys can wake up now. We are ready to start.

Just to talk you through a little bit how the flow of what we are going to do here today, we are going to start with a brief introduction of the measure from the Co-Chairs, the title, developer description.

We will have the developers give you guys an overview of the measure. There have been a few, I think, last-minute changes to the measures which are in your handouts. The developers will kind of be talking you through those and bringing your attention to specific aspects of the measure that you may need to be aware of as you dive into the criteria.

As you know, each of you has been assigned a piece of the criteria. We wanted to make sure that everyone was engaged. And so, in your PDF handout that we sent out via email -- and I am not sure if it is printed out, but definitely in the PDF, and we can bring it up on the screen -- we have made assignments to everyone for pieces of the measure. So, the Co-Chairs will be calling on you when we get to that portion of the measure to give your overview.

We will also be putting slides on the screen that will kind of cue you to some

of the discussion questions and things you should be highlighting during your presentation for each of the subcriteria. So, we will be here to help walk you through that.

We are going to be kind of hands-on with this first measure to make sure everyone kind of gets their feet wet and gets a feel for kind of the flow. And so, we will kind of introduce the measure, the criterion we are looking at.

We will have the lead discussant pick up and give a description summary of the submission relevant to your assignment, give a summary of the relevant Committee preliminary recommendations -- I'm sorry -- preliminary recommendations and comments, and highlight where you think there may need to be more discussion. Or where there is specific agreement, we will try not to spend a lot of time there, but really focus on the areas where there seems to be a lot more issues to dive into for that particular area.

And then, we will be asking you to summarize any of the public comments that were relevant to your specific criteria to make sure those are addressed and are included in the discussion.

Once we kind of have that overview, we will, then, open it up for Committee discussion. So, the lead discussant is really there to kind of jump off the discussion, highlight what we need to talk about, and then, we will open it up to the whole Committee to dive in and provide input. And then, David and Gene will help facilitate that discussion.

And then, when you guys get to a point where you feel like you have got everything out, we will have you vote on the criteria. One thing we have kind of added in, I guess, ad hoc on the fly, we have a number of people here who are in the room from the public, and we will give them an opportunity to provide comments before you vote, so that

you have all the input.

We want to try to have everything on the table. Once decisions are made, sometimes it is hard to go back. So, we want to make sure everyone has all the information before them as much as possible.

We will have you vote on the overall criteria, and depending on the results of that vote, we will move on to the next criteria, and we will just go systematically through each one like you did in your preliminary evaluations.

Once you vote, we will ask the Co-Chairs to kind of summarize the discussion to make sure that what we see reflected in the votes and the discussion that we heard actually kind of align, so that people understand that what they see in the votes is actually reflected by specific things that were raised as concerns or what have you.

So, does that jibe with everyone? Does anyone have questions about the flow of

how we are going to evaluate the measures?

Jack, you have this look on your face.

MEMBER NEEDLEMAN: No. No, I am fine.

(Laughter.)

MS. WILBON: I am like, is that good or bad? Sorry, not to call you out, but I want to make sure we are all okay.

MEMBER NEEDLEMAN: If I have a problem, Ashlie, you will know it.

(Laughter.)

MS. WILBON: Okay. I know. I know. Thank you.

Okay. Can you go to the next slide, Evan?

So, everyone should have a little pointer, an electronic voting device. As we get to the criteria, there will be a slide that comes up on the screen. We will probably do a test at some point. Yes, we should probably do a test to make sure it is working.

We will be asking you to vote. You will have 60 seconds to enter your vote. On the slide, it will tell you what number you hit to correspond to each of the rating scales. The votes are anonymous. We don't know necessarily right now you are voting. So, you don't have to feel scared about that.

And then, once all the votes are collected, it will display on the screen, so everyone in the room will know what the outcome of the votes was. And we will have either staff or Co-Chairs recite what that is, so everyone on the phone and in the room knows what the votes are.

The voting device that collects your votes is actually on this laptop right here where Evan is sitting. So, we will be asking you to point in that direction. Okay?

Any questions about that?

(No response.)

Okay. I think that is it. We won't go over this.

Okay. So, this is getting into the first criteria. And we will have, I guess, one of the Co-Chairs introduce the first measure.

CO-CHAIR PENSON: Sure. Sure.

Why don't we get started there with the first measure, which is 2158, which is Medicare Spending Per Beneficiary? I will just sort of summarize this as I understood it.

Very quickly, I think it is, for these sorts of measures, relatively straightforward. And that is an important modifier right there.

So, basically, what this measure attempts to do is at the hospital level assign cost to an inpatient episode. It starts three days before the admission for the inpatient episode and follows out to 30 days after the inpatient episode.

It captures all A and B charges. Basically, once it does that, it basically

generates an O-to-E ratio where the expected ratio is generated by the MS-DRG, and then, that is modified against a national median, if I understand it correctly. And if I don't, I apologize. And that comes up with the final measure, with 1 being, as I understood it, fairly reflective of the median and, obviously, above 1 and below 1 going in either direction.

So, I think that sort of summarizes what it is at. The goal, as I understood it, was to sort of look at the hospital and try to improve coordination of care. As often, as they say in the measure, hospitals will discharge patients perhaps too early, and that can lead to readmissions. And by looking at the care around the hospitalization and counting resource use, we may be able to get towards better value by looking at the cost on either side.

So, I think that is fairly the overview of the measure. I am sure we will

get much more into the weeds as we go forward, and it will turn out that everything I just said was wrong.

I'm sorry, and the developer, if I am not mistaken, is CMS.

MS. TIGHE: Yes. If someone from that developer group wants to come and present the slides that you all put together and introduce the measure?

MR. ZAIDI: Good morning.

My name is Sajid Zaidi. I am with Acumen LLC. We are co-measure developers of the MSPB measure, under contract with CMS.

So, I just wanted to give a brief overview of the MSPB measure. And then, we would be happy to take any questions you might have.

So, the MSPB measure was actually mentioned by name in the Affordable Care Act as part of the hospital value-based purchasing initiative. CMS was required to develop this measure.

It measures total Medicare-allowed cost for hospitalization episodes. The hospitalization episode includes all Medicare Part A and B claims from three days prior to the index admission date to 30 days after the hospital discharge date. So, the only thing that is really not included is prescription drug claims, Part D.

It is an all-cause measure. It includes all conditions.

The measure is payment-standardized and risk-adjusted to allow for a comparison across all hospitals in the country, and it applies specifically to Medicare fee-for-service beneficiaries who are discharge during the period of performance from IPPS hospitals. So, these are hospitals that are paid under the Inpatient Prospective Payment System.

And so, the only main exclusion there are critical access hospitals and Maryland hospitals, which operate under a

waiver.

This measure was developed specifically for Medicare fee-for-service beneficiaries. We utilized the official CMS price standardization methodology, which has been adopted by the Institute of Medicine for their ongoing work on geographic variation.

The risk-adjustment model is an augmented CMS HCC model. The main difference between the HCC model that we all know and what we use is that we have added the MS-DRG as a risk adjuster.

And there are no similar NQF-endorsed measures to profile resource use around hospitalizations for Medicare beneficiaries.

So, as far the importance and usability of the measure, it is intended to be used alongside quality measures. So, when used in conjunction with other quality measures that CMS has developed, the MSPB measure allows one to identify hospitals that

provide high-quality care at a lower cost to Medicare.

So, for example, right now, it is already a part of the Inpatient Quality Reporting Program. It is publicly reported on Hospital Compare. You can go online right now and see MSPB scores for most hospitals in the country. And starting in fiscal year 2015, it will be a component of the Hospital Value-Based Purchasing Payment Adjustment, in conjunction with process-of-care measures, patient-experience measures, and outcomes measures.

So, the intent of the measure is to incentivize hospitals to reduce Medicare spending and delivery system fragmentation by improving coordination with post-acute providers to reduce, for example, the likelihood of hospital readmissions, unnecessary inpatient services such as multiple CT scans when they are not warranted; by reducing unnecessary post-acute care

services, and shifting post-acute care from intensive settings, such as the skilled nursing facilities, to less intensive settings, such as home health or other outpatient settings.

One thing I want to emphasize is that the MSPB measure, we have provided an unparalleled level of transparency to hospitals. We send each hospital, through the QualityNet website, three very detailed files that allow them to deconstruct the measure.

So, the first file that we send them is called an Index Admission File. This lists every single admission that the hospital had for Medicare fee-for-service beneficiaries during the period of performance. It includes the standardized cost for each admission and the episode as well as diagnosis codes, length of stay, and the major diagnostic category that the admission falls into.

And in this file, we also provide an indicator for whether or not the admission

was excluded, and there are various exclusion criteria that we will get into later.

The second file that we provide is an Episodes File. This is really the crux of our transparency efforts. This file lists for each service category, such as skilled nursing, physician cost, hospice cost, or home health cost, it lists the amount of cost in each of these settings. So, a hospital can really break down for every one of their episodes what are the components of the episode.

And this file also lists the top five providers for care provided in each of these settings. So, it really makes this information actionable, if you know exactly which providers are providing skilled nursing services or home health services to your patients.

And the last file is a Beneficiary Risk Score File, which provides all of the risk adjusters used for each index admission

as well as the coefficients on each of those risk adjusters. So, if the hospital wanted to and had the analytic capability to do so, they could reconstruct from scratch the risk adjustment part of the measure.

And the last slide is sort of a nitty-gritty list of all the steps in the construction of the measure. I will just go through it really quickly.

So, the first step is we exclude inpatient admissions that contain a transfer or where the beneficiary died during the episode or did not have continuous fee-for-service enrollment.

The next step is to construct the MSPB episode by including all standardized Part A and B claims from three days prior to admission to 30 days after discharge. And one important thing to note is that an admission can only start an episode if it is not already part of another episode. So, you can't have an admission as part of the 30-day window of

another episode and also start its own episode.

The next step is the risk-adjustment step. We calculated expected spending, adjusting for case-mix variation using the HCC model and the MS-DRG of the index admission. And the risk adjustment is actually done within a major diagnostic category. So, we run multiple regressions, one for each MDC.

The next step is excluding outlier episodes. These are episodes that had a residual which fell above the 99th percentile or below the 1st percentile. So, these are episodes that had predicted spending which was either far below or far above the actual observed spending.

And then, finally, the hospital's MSPB amount is calculated as an efficiency ratio, which is the ratio of the observed-to-expected spending over all of their episodes.

And the final measure that is

reported is the ratio of this MSPB amount to the median hospital MSPB amount. So, the interpretation is, if the measure is above 1, you cost more than the median hospital; if it is below 1, then you cost less than the median hospital.

So, I think that summarizes our brief description of the measure. I would be happy to take any questions, either now or throughout the discussion.

CO-CHAIR PENSON: I would suggest that we hold the questions as we go through the individual criteria, and that way, if people have specific questions.

So, I think, without any further ado, we should actually get into the various criteria. And everyone had their song. And so, I am going to ask Larry and Lina to start the discussion around the importance of this measure.

MEMBER WALKER: How do you want to do it, Larry?

MEMBER BECKER: You go ahead.

(Laughter.)

CO-CHAIR PENSON: So, Lina, I think that means it is on you. Okay?

(Laughter.)

MEMBER WALKER: Right. Very well. Very well. I am glad I did my homework then.

So, I am going through the measures 1a through c. And not having done this before, I am just going to briefly summarize what the scores are, summarize the comments, what I read from the public comment, and then, if there is anything that I miss, just jump in and remind me to fill in the blanks.

So, for the first measure, 1a, impact, the summary score is 20 high, 1 moderate.

And I am not going to repeat the question. But the summary of the comments were, generally, most people felt that it scored it a high and had positive comments.

There were some comments that this measure did not demonstrate variation in total resource used by the hospital and, in fact, that it did not demonstrate performance gap, although one commenter noted that the evidence is available in the literature.

Should I just proceed to 1b? So, 1b is the opportunity for improvement. And the summary scores are 15 high, 3 moderate, 2 lows, and 1 insufficient evidence.

And the summary discussions were mostly around the -- hang on a second -- around the disparities issues. One commenter thought that the measure did not address how it could be, wasn't sure how it could be used to address or improve disparities. And another felt that the measure should adjust for socioeconomic characteristics. But, by and large, a lot of the comments were -- well, it is hard to tell what were the comments for 1b.

For 1c, the category was measure

intent. It seems like there were a lot of comments around this. Some of the comments were: does a lower score on MSPB reflect improvement? And that is not clear.

Similarly, some felt that the MSPB measure does not equate with efficiency, which was how the developer had described that measure.

Many felt that it was more of a resource use measure, although one commenter noted that it really wasn't a total resource use measure, but more a measure of Medicare spending.

A few felt that the measure doesn't demonstrate, that the developers do not demonstrate how the measure would improve care coordination or achieve some of the stated goals.

A few felt that the measure was not specific or targeted enough. For instance, some of the comments were that it did not capture discharge planning, care

coordination, nursing, and other spending that could reduce readmission rates, which would be a target for improving care coordination.

One commenter felt that the average measure might not be able to achieve the stated outcome rather than a condition-specific measure.

And a comment that the measure does not capture appropriateness or effectiveness of care.

There were some questions from some of the reviewers on whether certain service categories were included in the measure. So, one felt that it was not transparent, what was included was not transparent.

And then, there were questions about whether transportation, home health, physicians, drug costs were included -- whether transportation, home health, or physician costs were included the measure. And then, a couple of commenters felt that

drug cost should be part of the measure and, also, in-kind services should be part of the measure.

And then, finally, many felt that -- but, in general, though, I would say that many felt that the episode-based measure was directional correct. There was one comment that the three-day pre-admission measure captures spending that is outside of the hospital's control.

There was some concern about unintended consequences, potentially stinting on post-acute care.

And then, another note that the name should reflect the unit of analysis, which is really inpatient hospital episodes.

And then, I think I didn't say this, but the summary score for 1c measure intent was 12 high, 6 moderate, and 3 low.

There were a couple of comments from the public on this measure, but they were more general comment that applied to this

measure and the second measure that we will be evaluating today.

And some of them were about shared accountability and attribution. Some felt that socioeconomic status should be considered within the measure. And this came up because of the disparities discussion.

Some felt that there are variations in spending that is due to patient or community characteristics that may not be reflected in race and ethnicity. Maybe this is more of a risk-adjustment section. Okay, so I am going to skip that.

So, that is all for now.

CO-CHAIR PENSON: Great. Thank you.

MS. WILBON: I am sorry. I know Nancy Garrett I think had 2c, which was under disparities, and we are kind of pulling you into this discussion. So, I just wanted to make sure that you had an opportunity, too.

MEMBER GARRETT: Do you want me to

do that now before we open it up for comments or?

MS. WILBON: Yes, if you had any kind of assessments on preliminarily what was submitted.

MEMBER GARRETT: Okay. I have a few slides. Evan, would you mind pulling those up?

So, in terms of the disparities information, it is a complex issue. So, I pulled together some of the information, just to get us all on the same page before we have the discussion.

So, on the next slide, the first question that I was asked to talk about is: were disparities identified in the demonstration of importance in the submission?

And so, if you could just make that bigger? Thank you. Perfect.

So, the developer cited a number of specific public studies. I just pulled out the key points from these, and the citations

are in your materials.

But, basically, in terms of SES, socioeconomic status, low-income Medicare beneficiaries are more likely to use inpatient services and home health services, suggesting that low income could be correlated with higher cost, although the home health services, it wasn't clear to me whether that would be higher or lower cost. So, I don't know all the details on that study.

And then, in terms of race, a number of different studies were cited; for example, end-of-life hospital care, where the study found that Blacks and Hispanics were more likely to be admitted to the ICU and receive more intensive services than Whites. In another study, Blacks had more money spent on them than Whites, but had less effective interventions. Another study, Blacks were more likely to be readmitted, a study of readmissions for AMI, CHF, and pneumonia, than Whites. And then, patients from minorities,

at the hospital level, patients from the hospitals that had a higher percentage of minorities had higher readmission rates. So, there is evidence in the literature that disparities are related to cost, I think is the overall summary.

And then, in the next slide, another question for us to consider is: were disparities identified in the demonstration of importance in the submission? And here, in terms of the actual analysis itself, not just from the literature, but the analysis itself of this data, episodes for dual-eligible beneficiaries cost more, $859 more. But, then, at the hospital level, when they modeled this out, the results were similar in hospitals that had higher shares of dual-eligibles and higher just percentages.

So, it was interesting that at the beneficiary level there was a difference in cost. But, then, when they modeled it out at the hospital level, they didn't see that

difference.

Now a key point here is that the percent of dual-eligibles and the just percentages, note those are proxy measures for SES. And so, are those really capturing the underlying differences is a key question. And then, you have the different units of analysis, the individual in the hospital. But that contrast there is an interesting thing for us to discuss.

And then, on the next slide, now you get into the question of, okay, well, if there are disparities, what do we do about them? And so, that starts to get into the risk-adjustment discussion. And so, I will just talk about that a little bit, and then, we will talk about it more in the risk-adjustment section.

And so, at this point, the measure isn't including these variables in the risk adjustment or stratifying after the fact. And in terms of the risk adjustment, the rationale

is, as we heard already, that NQF has a position on not adjusting for socioeconomic factors.

And then, secondly, the measure developers said, "Well, but let's try it out." And so, they actually tried excluding dual-eligibles or adding them as a risk-adjustment factor. And that didn't result in major changes in the results or in the risk-adjustment models. So, they actually went ahead and tested it and didn't see major differences, which is interesting. So, another point for us to talk about.

And then, finally, the last slide, some of the comments that people submitted on this question of disparities and, then, what to do about it. One person said, "This could be a validity issue. So, why is it that we are not seeing correlation with proxies for disparities? Given that we do have this strong correlation on individual level, is that a validity question for the measure?"

Another comment, "Validity of the measure has not been fully established due to lack of adjustments for age and race." So, even though there is a policy not to adjust, there are clinical factors that are present at the start of care and should be considered. So, I think there is even discussion here about the policy and whether it makes sense in this case or not.

Another comment strongly encourages the Committee to discuss the viability of addressing socioeconomic status within the measure.

And then, another person said, "Yes, we see disparities, but we think the developers have a rationale for not stratifying."

So, there are some mixed comments there. That just kind of kicks off, I think, the discussion here.

CO-CHAIR PENSON: Great. Thank you.

So, I was just talking to Ashlie. What I think we should do is go through the three importance criteria, 1a, 1b, and 1c. And as we go through each of them, we will vote on each of them in turn. And there are questions. You will see the slide will come back with the questions we are supposed to address. Nancy's discussion probably wants to come up in the 1b.

To reflect back on what Lina said, to sort of summarize, when we are talking about 1a, what I heard -- and correct me if I am wrong -- that there was general agreement around 1a that the importance was generally high, but there were really no issues around 1b. And given what Nancy said, it sounds like there may be significantly more issues because of the way it is parsed out. And perhaps we should try to go in order. But I think we will have some longer discussions about 1c because I think there were some questions about intent.

So, why don't we try to just go through each of these in turn? What I would say we should do is we can just start and open up the floor to discussion. And if you have questions amongst Committee members, that is great. And also, with the developer here, they can address them.

So, the first thing we looked at was 1a, which is the priority of the nature. Is this something that the specific national goal of priority identified by HHS or the National Priorities Partnership, I mean by NQF? Or this is an high-impact aspect of healthcare.

And like I said, I got the impression that most people felt that this was important.

If we can go back one slide before that? There we go.

So, the discussion points really to think about are: are there large numbers affected by the measure? Does the measure

demonstrate variation in resource use or overall performance? And are there patient/societal consequences of high or low resource use?

So, I would open it up to discussion on the floor. Just put up your tent if you want to say something.

(No response.)

Very quiet. Okay.

There were some people who at least started with moderate. I don't think there were any lows here. If that person who is -- I don't want to call anybody out -- but does anyone feel any concerns or questions here?

(No response.)

Okay. Well, hearing none, I would say we probably should vote on it.

Do you want to do a test first, Evan, or just go right and vote and see if it works?

MR. WILLIAMSON: We don't have to

get a test set up, but we can use this slide to test it. And if it works, great, then we will keep the result.

So, I just want to make sure everybody points their clicker at this little dongle here on the computer. We will start here. You have a minute to vote, and your minute begins now.

(Vote taken.)

CO-CHAIR PENSON: I went to a meeting once, and they had the Jeopardy music on when they were doing this.

(Laughter.)

How do we know that we pointed at the right thing and our vote gets counted? I feel like I have a hanging chad here.

(Laughter.)

DR. BURSTIN: Even if you push twice, it only gets read once. So, it is not like people can vote multiple times. And we will get to see a count of how many have voted. And then, we will get a sense if it is

working.

MR. WILLIAMSON: I counted 25 earlier. So, I just want to make sure everybody points. There we go. All right, we are at 25.

CO-CHAIR PENSON: Okay. So, we have 23 high and 2 moderate. So, the vast majority of folks think it is high, and I think that is good.

So, why don't we go on to 1b, and I expect there will be more discussion here. So, just to remind you, this is demonstration of resource use or cost problems and opportunity for improvement. And I want you to note that it includes a comment about data demonstrating variation in delivery of care across providers and/or population groups, which brings in that disparities issue that Nancy was talking about.

And we will come back to this later when we go to component 2, but I think that we want to think about this here as well.

Because if you feel that it is inherently flawed in picking up disparities, then I think it has to be discussed here. If it is more of a nuance, perhaps it should go on 2.

Some of the discussion points that you can see include: do the data demonstrate distribution of performance scores? Is the number representative of the entities? Do the data show disparities in the use of resources across care across certain populations? And then, size of the population at risk.

So, folks who want to start a discussion -- and there were some people here. I think there was a low here, if I remember correctly. So, I wonder, folks who have concerned, go ahead and put your tents up if you want to talk.

Jack, why don't you start us off?

MEMBER NEEDLEMAN: Okay. I think there are several issues here, not about the overall importance. After all, Congress asked for it. That is prima facie it is an

important measure.

The question is whether it actually delivers on what it promises to do. And I think there are three issues here that are all related.

One is it is accurately labeled as Medicare spending. But all through the discussion there is a discussion of cost, that it is a cost measure. It is not a cost measure; it is an expenditure measure.

And indeed, given that they are standardized, it is really a dollar-weighted unit-of-service measure, aggregate numbers of units of service. So, we get all this resource cost compression in every one of those units. So, if there are differences in how much hospitals are spending, it is invisible to this measure. If there are differences in what a physician's office is spending, if it isn't showing up on a bill, it is missing from this measure. It disappears.

So, we have got that issue in

terms of understanding. If we are trying to say Medicare expenditures, it is not bad. If we are trying to understand resource use or cost, it doesn't fully satisfy, it doesn't meet that concern.

When I think about what will drive this measure up, what are those cases at the high end? Given you start with a hospitalization in the middle, everybody has one and we have controlled for the DRG, so everybody's standardized hospital cost is already in there as the DRG, what is going to drive the differences and what is going to drive somebody high is readmission and SNF use.

Because other things are much smaller in magnitude. More doctor visits, not going to potentially make that much difference.

So, we have got the question of whether a measure that is fundamentally measuring differences in readmission and SNF

use -- and I may be oversimplifying and I would like to hear the Acumen people talk about what they see as driving up the high-end volume here, not in terms of dollars, but in terms of what units of service are driving that up.

But, basically, is a measure that basically captures differences in SNF use and readmissions an adequate measure to deal with the importance of measuring Medicare spending per beneficiary per hospitalization episode?

CO-CHAIR PENSON: So, I wonder if the developer could comment a little bit on that, only because I think Dr. Needleman's point is basically that this is really going to be driven by a few key categories in utilization. I don't know if you have found that or if you believe that is true or not.

MR. ZAIDI: So, we did look at this. We did a study decomposing the variants in episode cost across all hospitals. And you're right that most of the variance is

accounted for by broadly-defined post-acute care, which is SNF, home health use, and, of course, readmissions. So, those three categories do account for most of the variance across episodes, but I think it is important to remember that those are the high-impact categories where Medicare spends the most money surrounding hospitalizations.

CO-CHAIR PENSON: So, Jack, what do you think about that?

MEMBER NEEDLEMAN: So, then, the question is, are variations in that driven by patient-level factors that are not fully captured in the risk adjuster or are they, in fact, differences in resource use that reflect choices in care that are in some sense under the control of or influenced by the providers?

CO-CHAIR PENSON: I think that is a very reasonable point. I am curious to get other folks' thoughts on that.

Bill, you have your tent up.

MEMBER WEINTRAUB: On each of the

points that Jack just raised, first, is this a measure of cost? Well, yes and no. What we are interested in, in principle, is total societal cost, but we can never get there. So, we use proxies.

I agree that, for many economic studies, I prefer using resource use of a hospital and trying to get at true hospital costs. But what Medicare is spending is a fair measure of cost. After all, we are paying our taxes to the federal government and they are spending those dollars. Some economists will say, "Well, that is just transfer of dollars. It doesn't mean anything." But others would disagree.

I think this is actually a fair measure of cost, seeing as there is no perfect way. I agree, though, that what we are looking at primarily, because, after all, the payments for the initial hospitalization are DRG-based, what we are primarily looking at are SNF use and readmissions.

And there, the real question is, is it fair to attribute these resources downstream to the initial hospitalization? And there is lots of noise because sometimes yes, sometimes the hospitals fail to do things and patients are readmitted. And that is sort of a subtext that runs throughout the measure as it is presented to us. But sometimes they are not. And so, there is a lot of noise.

Nonetheless, this is going to be used. It is already being used, and hospitals are very much tuning into the idea of taking total risk, including what is coming downstream.

CO-CHAIR PENSON: Joe?

MEMBER STEPHANSKY: Not all hospitals are ready to take on that kind of insurance risk.

(Laughter.)

Thank you, Jack, for your comments because I buy into all of that.

As far as Acumen, I know this

particular application for the measure endorsement doesn't reflect the work that you did for the Institute of Medicine, but that was essentially the same kind of finding, that it was the post-acute care, the differences in that, that were driving much of the variation. And I think home health played a bigger part in the work you did for the Institute of Medicine as far as the volume.

I would just like to remind people that, when a discharge occurs at a hospital, the discharge planner does not have the ability to say, "Okay, you need home health. You are going to this home health agency." There is a lack of accountability there. I mean, the patient gets to make some choices, not the hospital discharge planner, not the hospital itself. And we need to bear that in mind when we are trying to link accountability for the costs to Medicare after a discharge.

CO-CHAIR PENSON: Thanks, Joe.

David?

MEMBER REDFEARN: I think Jack made a really nice distinction. If you are looking at the cost downstream post-discharge, presumably, some of those costs are due to things that the hospital may have done or may not have done.

But the other part of it, it could be due to patient characteristics and unmeasured patient characteristics and incomplete risk adjustment of the patient characteristics.

And to jump ahead to one of the comments I had about the risk methodology, the HCC model, you got the MS-DRG, so that is going to be pretty good characteristic of what is going on in the hospital. But the HCC is not a great risk model. It is a real subset of a full model looking at a lot of things.

The other thing that really concerns me about this is they are capturing HCCs based on 90 days pre-admission. All of these models are designed for 12 months of

data. There is abundant evidence, looking at these kinds of risk models, that the accuracy of those models deteriorates dramatically if you have less than 12 months of data that go into them. In fact, it is the slope drops off dramatically at about seven months, and you are getting down to three months of data here.

So, this model is not going to adjust accurately for those patient characteristics, which may influence the post-discharge costs. And that concerns me a lot about this methodology.

CO-CHAIR PENSON: I want to raise a point as I understand it, and the NQF staff can correct me if I am wrong, which I may be.

But we will get an opportunity to go into attribution later and risk adjustment later. I think these points about attribution that were made before and risk adjustment are reasonable to raise here, but I think we want to look at it from a sort of 30,000-foot view. So, in other words, is it so flawed that you

have no chance for opportunity for improvement? Or is there still some opportunity for improvement, but there are flaws here?

And one of the things I heard from Nancy's discussion -- I think, Bill, that you are raising it as well -- is, when you look at this, are there patient-level characteristics that are so out of control of the hospital that they make the entire methodology invalid?

So, I am curious, as I hear people saying they have concerns about the risk adjustment, about the attribution, is it sort of a non-starter? That is what I think we are hearing. Correct me if I wrong about that, Ashlie.

MS. WILBON: Yes, David, you are right. And I just want to remind people about what we are looking for. For this 1b criteria, that the developer has demonstrated that there are resource use or cost problems in general, that this measure is needed, and

that there is opportunity for improvement. And by showing that there is opportunity for improvement, they would have data that there are gaps in performance across providers or gaps in care delivery for a certain population. So, really focused on whether or not they have demonstrated there is a need for this measure because there needs to be improvement in either how costs or resources are used or how care is delivered to certain populations.

So, just to kind of refocus the discussion, because I think the points you are bringing up are great, but will probably go in scientific acceptability.

CO-CHAIR PENSON: So, Lina, go ahead.

MEMBER WALKER: Is this the point where we talk about the disparities? Can I talk about 2c?

I raised the concern that the measure didn't vary with proxies for

disparities. To me, that seemed to be a significant limitation of the measure, given that we know that there are significant variations in spending.

And it raised the question about the validity of the measure, and I don't want to go into that now. But I would like to hear a little bit more from the developer about what they thought could be driving the lack of variation and whether there were opportunities to use different types of proxies because, clearly, the ones that you used were inadequate, and perhaps they were just weak measures of disparities.

MR. ZAIDI: So, as the measure is reported on Hospital Compare, there is no stratification for socioeconomic status. But for the submission form, we did look into a few measures of socioeconomic status. So, it is hard to say exactly why there wasn't as much variability or as much of a disparity as you would expect. But I think one thing is

that we are conditioning on the reason for admission for the initial index admission. So, if you would expect that dual-eligibles are more likely to be admitted for more severe care, we are controlling for that by controlling for the MS-DRG of the index admission.

And I would say that our analysis showed that Medicaid beneficiaries had about $1,000 more expensive episodes on average. That didn't translate into a clear pattern between hospital performance. I think that would take further investigation to figure out why the per-episode variation didn't translate into between-hospital variation.

MS. SPALDING BUSH: Can I just add something? One of the things, this measure also went through notice and comment, rulemaking in the FY 2012 IPPS rule. And one of the comments that we received with regard to socioeconomic status was the concern that Medicare beneficiaries of lower socioeconomic

status would have multiple chronic conditions that we needed to account for. I think that perhaps that we capture the HCCs for every claim that is billed in the 90 days preceding the admission may help to have mitigated some of that concern.

CO-CHAIR PENSON: Thanks.

Martin?

MEMBER MARCINIAK: So, the question I have actually comes back to 2b. It was something that you touched on in your slide there.

And so, given that this gives the opportunity to drop down to the provider level and to start, quote, "changing behaviors," I would like to hear a little bit about how or what evidence you have to show at this point that it is actually changing behaviors.

MS. SPALDING BUSH: I think the measure is fairly new. So, we haven't seen a lot of change in behavior, but I don't think we have any analyses that demonstrate that.

MR. ZAIDI: No. So, I think it has been posted online for maybe a year now.

MEMBER MARCINIAK: Nothing that is sort of preliminary out there that is showing that, when people have actually used this, it has actually made a difference in terms of how they are actually practicing at the hospital level?

MR. ZAIDI: Nothing that specific that would say that there is change in behavior that is attributable to the MSPB measure.

CO-CHAIR PENSON: Matthew?

MEMBER McHUGH: I think this is maybe related to what Martin said and, in part, is asking for some clarification about maybe what the MAP is doing.

But I am having a little bit of difficulty with 1b. If we are asking about whether there is distribution in performance, which I think basically there is, but if we think about what an organization is going to

do with that information, what does improvement mean if we can't think about it in relation to quality or efficiency? I think that that makes it difficult to evaluate.

MEMBER MARCINIAK: Just to bridge off of Matthew on that, I mean, that is one of the challenges I actually had when I was reading this. The context actually matters. And so, to put resource there without context, you know, frankly, I struggle with that because you are saying that you are looking at resource and you are looking at a hospital. And then, it becomes the bridge to something else, something that is not entirely clear yet in terms of what the quality measures are because it is, in fact, it is what you are trying to drive to.

CO-CHAIR PENSON: I do wonder, getting back to some of the other slides, about the building-blocks approach because it is going to be tough for us to sort of make that next jump with both of these. I can

remember in Phase 1 that was a big issue. So, you sort of have to suspend disbelief on that piece and trust that CMS ultimately will develop value measures, but I think it is a well-taken point. I don't want to minimize it, but it will kill us if we get stuck there.

Daniel?

MR. WOLFSON: I want to reflect on your opening comment about the purpose of the measure. And I am getting more concerned about that.

You said it was about care coordination. And then, we have heard that, really, the measure is very dependent on use of skilled nursing facilities and readmissions. So, are we really, really wanting to measure care coordination? And if we really are measuring care coordination, would this be the measure that we would use?

So, I am getting a little confused about the intent of the measure. Frankly, I don't think that is true. I think it is about

cost, and it is not about care coordination. I think we would be measuring it in a much different way.

CO-CHAIR PENSON: So, don't kill the messenger.

(Laughter.)

No, that's okay.

And that may also be the way I interpreted it, but I will turn that over to the measure developers. It may have been that I misspoke, Daniel.

MS. SPALDING BUSH: Yes, I think that the care coordination piece fits into the measure when we talk about how could a hospital potentially improve their performance on the measure. But I think you are right that it is a measure of cost, of resource use. So, I think you are correct. And I think care coordination plays in when we talk about what could a hospital do.

And as Sajid mentioned when he was going through the slides, we provide them with

a vast amount of information. They can see for each category what are their most expensive service delivery types within their episodes. And they actually get the names of those providers if they have had an interest in trying to coordinate care with them, to see whether they could reduce duplicative services, unnecessary services, repeat tests, that type of thing. The hospitals are given all the information that they need to try to do that.

But you are right, it is not intended to be specifically a care coordination measure.

CO-CHAIR PENSON: So, if I could reflect back on this, because it helps me with my understanding as well, while you do feel that there may be some help with care coordination, it is clearly not a care coordination measure and you are going to report out to hospitals information of the individual provider and utilization, say, of

imaging within the hospital admission, et cetera. So that hospitals get that information for improvement?

MS. SPALDING BUSH: That is correct. And I am looking at Sajid because he knows better what it is, actually, they we provide in those hospital reports.

MR. ZAIDI: I don't think the hospital report breaks out imaging specifically as a category, but it will break out physician services during the index admission. So, you will be able to see all the professional fees that were incurred during the index admission.

CO-CHAIR PENSON: Okay. So, I am sorry. Sorry, David, go ahead.

MEMBER GIFFORD: Helen may kill me for this.

In the past on some of these panels, when measures have to be interpreted within the context of other measures, we have recommended they be paired with other

measures. And that is what I had when I was reading this through, that by itself this measure sort of is loss in context and meaning. We may want to think about saying that, if it is used, it has to be paired with some other type of measures that are out there to link at that issue, to get at the issue Martin was bringing up.

DR. BURSTIN: I won't kill you.

But this is an issue that has been longstanding. In fact, the NQF framework on episodes very clearly said resource use measures should not be used in isolation. And there was a recommendation, I think pretty clearly. And again, I think it is reasonable to -- I think we iterated that in the first report. I assume we do that again, that these measures should always be used in concert with quality measures to really get a full picture.

Anything you want to add on that, Taroon?

MR. AMIN: No.

DR. BURSTIN: Great.

CO-CHAIR PENSON: So, I think, David, would it be safe to say that you do have a specific quality measure you would pair with it in mind. Just you have to keep it in context. Okay?

So, Lina, I think you were next. And then, Tom and Brent.

MEMBER WALKER: Forgive for keeping repeating this question, but I just want to be very clear about what we are supposed to do here. So, I want to clarify, again, how we should evaluate this measure.

So, what I am hearing from developers is that we should think of this as a total cost measure, and you had used care coordination as an example. So, then, as we evaluate the broad applicability of this measure, some of the things that, for instance, I would want to consider would be how that measure could be used in multiple contexts. But, then, sort of the broadness of

the measure might produce limitations in the broader application; whereas, in the more targeted application that you had initially provided as an example, it might have been slightly more useful.

So, I just want to be clear about how I should be thinking and evaluating this measure.

MR. AMIN: Okay. So, that is a good point, Lina. And so, what we are specifically looking at right now is to look at the question of, in the context of our theoretical construct that we are operating under, which is that resource use measures are a building block to understanding efficiency and value, and where we agree that measures of resource use need to fit all these criteria -- important, scientifically-sound, usable, and feasible -- given that, which is the context in which we are having this discussion, so these points about the quality relationship is important. But, at this point in time, we are

looking at the scientific soundness of the measure of cost. Right now, we are looking at the question -- and again, at the 30,000-foot level -- does this measure give us an indication that there is opportunity for improvement in cost, meaning is there variation? Is there variation among providers? And is there variation among population groups?

These issues around disparities will clearly align with some of our conversations in the scientific acceptability portions around the appropriateness of the risk-adjustment model and the appropriateness or not of the stratification approach. And that discussion will occur at that point.

Thirdly, there is new information or new discussion, which I don't want to totally dismiss, but I think we will have to have it at some point, specifically in the use and usability discussion, which comes to the point of providers who are held accountable

using this measure, are they going to have enough information to facilitate improvement? And that will be the point to have that discussion.

And so, all of these are important. So, I would just encourage the Committee to consider the criteria as described, and through this first effort, try to focus the discussion there because it may be the cleanest way to have this discussion.

CO-CHAIR PENSON: Tom?

MEMBER TSANG: Yes. Sorry, I don't want to belabor this whole point about context. In my mind, I am thinking about putting this measure next to some of the 33 quality measures that an ACO or an MSSP program would be facing this year.

And so, I am still kind of unclear about how it would work next to perhaps a bucket of diabetes measures or a bucket of CHF measures. And so, perhaps CMS or Acumen can actually walk us through a use case. I know

we are talking about hypotheticals, but maybe that would help me think about the context.

CO-CHAIR PENSON: So, before you go ahead, I am going to ask the NQF staff, because this goes back to the concept of what is the mission of the Steering Committee versus the MAP, you know, the broadness versus the specific, and how far into the weeds you want us to get with this.

MR. AMIN: It is an important question. If we feel it is appropriate for CMS to respond to that, that is fair. The challenge is that there are two issues here. The first is that this Committee, again, is under the construct that the resource use measures need to be scientifically-sound and meet the endorsement criteria. And that is the function of this group at this moment.

Questions around how we move toward cost and quality linkages or how this is used in programs together is a discussion for later on this afternoon, pretty much at

the end of the day.

And the issues around how these would actually link up together in particular programs is really the purview of the MAP and is out of scope for this Committee.

So, what we are intended to do right now is to evaluate these measures against these criteria. So, I would encourage the Committee and the Chairs to limit the discussion to that task.

CO-CHAIR PENSON: So, I hear what you are saying, and I will respect it. I am going to ask Tom and others if they need to hear a little bit about a hypothetical specific program for context to assess 1b.

I mean, Tom, you raised the point. If you really feel you need it, then I think we should ask CMS. But I think you have to keep it in that broad view.

What do you think?

MEMBER TSANG: Well, I hate to them on the spot, but I think, for myself, for

hypothetical discussion, I think it would clarify just a little bit. I mean not a full-length discussion, but take a test scenario or a use case or something.

DR. BURSTIN: CMS is willing to respond.

CO-CHAIR PENSON: Okay. Joe is raising his hand. Maybe he --

MEMBER STEPHANSKY: Yes, I have an example, and this may seem kind of funny coming for me coming through thinking about the fairness of the measure.

But, right now, the Hospital Association, some University of Michigan faculty, and Blue Cross Blue Shield of Michigan are working on just this kind of application where we are using the commercial claims data from Blue Cross Blue Shield and a cost measure somewhat similar to this, cost to Blue Cross Blue Shield where we have a different standardization method, obviously, because of the range of contracting amounts.

But we are starting forward on looking at a total cost measure per episode using our own kind of episode groupers to examine just these questions because we see the variation; we know there are things we can do to reduce the variation, reduce the cost. We are linking them up to specific collaborative quality initiatives, say, in surgical infections or in urological surgeries, bariatric surgeries. We know there are ways that we can work together collaboratively to reduce costs while maintaining quality or increasing them.

It is a very similar concept to this measure. So, I have no problems at all calling the question. Let's just vote on this thing, say it is important, and worry about these issues in the usability area.

CO-CHAIR PENSON: So, we have a move to call the question, but I see there are still some tents up.

(Laughter.)

And I want to make sure everyone has a chance to say --

PARTICIPANT: I second that motion.

CO-CHAIR PENSON: Well, are you guys okay with that? Bill? Brent?

Use the microphone.

I think I am okay. I am okay unless someone needs another example, because I wasn't as concerned. But if others still are, all right.

Brent, are you okay calling the question, too?

MEMBER ASPLIN: That is what I was going to say. This is important. We have a lot of validity questions. So, let's get to them.

CO-CHAIR PENSON: All right. I sort of agree, but I want to make sure everyone has a chance to talk.

(Laughter.)

Very good. So, I think let's call

the question. Thank you, Joe.

So, go ahead and let's have a vote.

MR. WILLIAMSON: We will now vote on criteria 1b, the opportunity for improvement. You will have 60 seconds. You may begin now.

(Vote taken.)

And we have 12 high, 12 moderate, 1 low, and zero insufficient.

CO-CHAIR PENSON: Okay. Does this have to have all high or high/moderate okay?

DR. BURSTIN: High/moderate is okay.

CO-CHAIR PENSON: Okay. Great. Terrific.

So, here is what I would propose we do: apparently, folks on the phone can't hear me. Too close? Okay. Can you still hear me now? Can you hear me now?

(Laughter.)

It has been a long day already.

So, I think let's get to 1c. I think after 1c, which I think there will be some discussion, we will take a short break. So, I ask you guys to sort of work through this.

And I think that this is where people got into the issue of intent and linking to efficiency, or linking to quality, which I think we should be careful of. And I think Jack's point about this is really just resource use. That may be exactly what we are trying to do here.

So, the things you want to think about as we talk about is, is the intent of the resource use measure clearly described? Is the construction consistent with the conceptual construct? Do the resource use categories align with the intent of the measure? And as you heard, some are measured and some aren't. And are all the categories, types of resources, captured in the measure that you would expect, based on the intent of

the measure?

So, with that, I will open it up to discussion. Or, Joe, you can call the question right away, if you like. I like that idea.

(Laughter.)

Nancy, you have your tent up.

MEMBER GARRETT: So, I think in this section, and perhaps under the transparency section as well, to me, the name of the measure is a real problem. And so, I think maybe we are stuck here because this is written in the statute in this way.

But this is a measure of spending for beneficiary inpatient episode. And so, the fact that it says it is per beneficiary, to me, it is very confusing. As people are trying to get their arms around it, soon there are going to be a lot of cost measures, not just two. And so, I think it is really important that it is very clear what we are talking about here and who is being held

accountable and what the measure is.

So, I just wanted to kind of say that for the record on this section.

CO-CHAIR PENSON: Okay. Other comments?

Martin? And then, Jack.

MEMBER MARCINIAK: I would think that as well. I think that one of the problems that I had along with the transparency, frankly, was the listed categories. It was hard to discern exactly what those listed categories were within the measure.

We looked at the building blocks. We have spent a fair amount of time talking about those now. I think the blocks went resource efficiency and, then, value. Within the context of the document, it talks about this an efficiency measure and not a resource measure. So, it might be nice to clean or fix that as we go along as well.

CO-CHAIR PENSON: I wonder, for my

benefit and others, you sort of have alluded to it. Can you talk a little bit about the types of categories specifically which are captured?

MR. ZAIDI: Sure. So, the measure captures basically everything that Medicare pays for other than prescription drugs, everything other than Part D.

CO-CHAIR PENSON: I mean, how is it reported back? Just as the single number? Because I think that is, to me, some of the question here about the intent. And also, it goes back to 1b as well, the improvement issue.

So, I understand you are capturing everything but Part D. But, then, how is it sort of broken down when you do the reports, et cetera?

MR. ZAIDI: So, you are right that the headline number is just one index number, but in the hospital-specific reports and the data files that we provide to hospitals,

currently, it is broken down into seven categories, which are inpatient, hospice, SNF, home health, outpatient hospital, professional physician, and durable medical equipment. And those are further broken down by the three time periods of the measure, which is prior to admission, those three days prior to admission; during the index admission, and post-discharge.

Those documents, we can, in response to comments or suggestions, I think in future years we have the ability to change those data files that we provide to hospitals to make it more usable for hospitals.

CO-CHAIR PENSON: Jack?

MEMBER NEEDLEMAN: Two things. I just want to reemphasize that this is a measure of Medicare expenditures or Medicare costs. It is not a measure of resource use in the providers. The measure compresses, because of the standardization pricing, it compresses any variations in resource use

within providers, within billed served category.

And it excludes not only drugs, but anything else which is provided within a provider which is not a billable service.

So, those are two important limitations. I don't think the documentation of it fully captures those limitations. And in that sense, I am unhappy with the documentation.

Obviously, one of the big categories of excluded expenditures in there for resources that are used is drugs. So, do you have any sense of how much drug costs are? If they were included, how much they would be adding to this and how much variation there is in that? Is it an exclusion we ought to be worrying about if the intent is to measure Medicare expenditures and use the billing categories as a proxy for cost or resource use?

MR. ZAIDI: So, there were a

couple of concerns with including Part D. The biggest one is that only about 60 percent of fee-for-service beneficiaries have Part D coverage. So, we would be missing a large number of beneficiaries, and that would reduce the number of hospitals who would meet the minimum episode criteria to be profiled. So, there is a significant missing data concern with Part D.

The other concern is that the prices paid for drugs vary by plan. So, the actual spending varies a lot from plan to plan because they negotiate the prices for those drugs.

Kim, do you have anything to add on that?

MS. SPALDING BUSH: No.

MR. ZAIDI: But we haven't done a specific analysis on the level of variation in Part D spending.

MEMBER WALKER: I would have to clarify what you just said. Now Part B drugs

are included?

MR. ZAIDI: Part B drugs are included. Part D drugs are not.

CO-CHAIR PENSON: So, I just want to reflect back to Jack because I am hearing --

MEMBER NEEDLEMAN: I'm sorry. And Part A drugs are included, too.

MR. ZAIDI: Well, not specifically because I think those are bundled into the --

MEMBER NEEDLEMAN: That means they are included.

MR. ZAIDI: Yes.

CO-CHAIR PENSON: So, I want to just funnel back to Jack because I am hearing your concerns, and I am curious if this is, for lack of a better way to put it, a deal-breaker for you or just of moderate concern for you.

MEMBER NEEDLEMAN: A good question. It is probably a moderate concern. I think it is important in terms it will come

back in the usability discussion. So, I just don't like the mislabeling and I don't continuing to hear people talk about this as a cost measure. It is a Medicare expenditure measure, and that is different.

CO-CHAIR PENSON: Okay.

David?

MEMBER GIFFORD: I will go a little farther than Jack. I agree with his points, but, to me, I have concerns with how it is described as it is incentivizing hospitals to coordinate care and reduce unnecessary utilization. There are so many other ways to move this metric. I think it mislabels the intent and it mislabels how the results will be interpreted and used.

And so, I really have concerns how it is written here. Other issues with the measure, but I think the way it is written suggests the use and interpretation of the data that is very misleading and wrong. And I don't support it that way.

CO-CHAIR PENSON: So, what I am hearing from you, David, is actually a serious concern about the intent of the measure.

MEMBER GIFFORD: I am okay with the using of the measure and going out there. It is how it is written here. There are so many other ways to get and move this metric besides just incentivizing care coordination and unnecessary utilization.

And I think that we have heard a lot of the comments around here today. If we endorse it the way it is written, we would suggest that this measure is reflecting and changes are reflecting related to that. And I don't think -- all the comments we have heard say that there are a lot of other things that are contributing to the variation in this measure.

CO-CHAIR PENSON: So, I just want to reflect back because I think it is important because I am hearing very serious concern. I think it is in the right place

when we are talking about the intent of the measure. So, just to reflect back, I am hearing very serious concerns from Dr. Gifford.

Lisa?

MEMBER LATTS: So, a couple of comments related to what has been said. I agree, and I think probably everybody is going to agree, that to use this as a, quote, "coordination-of-care measure" is really directionally the wrong way to go.

So, maybe in terms of sort of moving the conversation along, we can agree, as part of our comments back to the developer in terms of the next phase, that we would like the whole coordination-of-care piece to go away. Since there are many valuable things, I think, that this measure can be used for, I just happen to think that is not one of them.

So, maybe so we don't keep beating this to death, which is probably too late -- (laughter) -- we can give that back to the

developer.

Then, I have two more comments. One, I am actually not bothered at all by the sort of resource use at the provider level. I, frankly, in terms of this measure, don't care. What I care about is what it costs Medicare or what Medicare is expending on it.

If a provider within their own facility is more or less efficient in providing those services, to me, that is not what this measure is about, and it is their business, and they need to work that out. So, I am not bothered by that at all.

I am bothered by the pharmacy and the Part D, and I really wish -- I know that there is a significant percentage of beneficiaries that don't have Part D coverage, but I really wish there was a composite of this measure with Part D for those 60 percent, or whatever percentage it is that had Part D, because I am bothered by that.

CO-CHAIR PENSON: So, I think the

Part D issue, that is, for lack of a better way to put it, set in stone. In other words, you have developed the measure that way and it is going to be what it is going to be. And we have to decide if that is okay within the measure intent.

With regard to the coordination of care, I think perhaps I get to bear some blame for this for highlighting it in their submission. But I will ask the CMS folks, I mean, I am interpreting your comments as that is an example of one of the ways, but that was not your primary intent. Is that a fair statement?

MS. SPALDING BUSH: I think it is fair to say that coordination of care we view as one of the ways a hospital could potentially improve their performance on this measure, yes.

CO-CHAIR PENSON: But not the only way, and that was not your intent? You are not looking at this as a coordination-of-care

measure?

MS. SPALDING BUSH: Correct.

CO-CHAIR PENSON: Okay.

MS. SPALDING BUSH: And we think avoiding unnecessary utilization at the hospital level is certainly another way to go and improve performance on that measure.

CO-CHAIR PENSON: Larry?

MS. SPALDING BUSH: Provide good care during the hospitalization to avoid readmissions, all of these things that could be done, yes.

CO-CHAIR PENSON: Larry?

MEMBER BECKER: So, this is CMS data, all these pieces, and they are a big purchaser, certainly a lot larger than I am as Xerox, right? But this big purchaser is trying to identify unwanted variation in resource use, and it is trying to send a sentinel message to a portion of its supply chain as a message to look inwardly and make process improvements.

So, yes, some things aren't included, and we would like to have it all, and we would like to have all of the pieces. But there is a sentinel message, and there is a bigger picture, I think. And so, I think it is important to put this out there, to get people focused on looking at resource use.

CO-CHAIR PENSON: So, Daniel, you had your tent up and you took it down.

I just want to remind folks, if you want to be heard, just put up your tent.

MR. WOLFSON: I just wanted to say, David, that I don't think that you highlighted it. It was highlighted in the report. I mean, it clearly says care coordination is the purpose of this measure. And I just think that is a little offbase.

CO-CHAIR PENSON: I am just a dumb surgeon. There is no way I would have come up with that on my own. Okay?

(Laughter.)

All right. I think this is

another one we have sort of been around a fair amount. So, unless there are any other burning questions, I think would call the question on this one, to vote on measure intent as you see it there.

So, Evan, if you want to turn that on?

MR. WILLIAMSON: We will now vote on the measure intent. You will have 60 seconds.

(Vote taken.)

And we have 6 high, 16 moderate, 3 low, and zero insufficient.

CO-CHAIR PENSON: So, what we will do now, I am told, is we need to open things up for public comment from people in the room and on the phone. And once we do that, we will take an overall vote for category 1.

So, any public comment either on the phone or in the room?

(No response.)

THE OPERATOR: For the audience

over the phone to ask a question or make a comment, please press *1.

(No response.)

There are no questions.

CO-CHAIR PENSON: Great. Thank you.

So, what we will do now is we will do an overall vote for all three of these categories.

I will open the floor if there are any burning questions before I call this.

MR. WOLFSON: I have a burning comment.

(Laughter.)

I think what I have seen in the report that they have supplied is kind of a case study of a report that would be produced by this data. And it does get into a granular fashion. I don't know if a hospital would like it as such, but it does get into some more detail than just the high-level stuff that we have been generally looking at. So,

I think there is some hope that they can pinpoint areas for improvement.

CO-CHAIR PENSON: There is always hope.

(Laughter.)

Other comments?

(No response.)

All right. Hearing none --

MR. WILLIAMSON: We will now vote on the overall importance to measure and report. You will have 60 seconds. Begin now.

(Vote taken.)

And we have 8 high, 15 moderate, 1 low, and zero insufficient. The measure passes overall importance.

CO-CHAIR PENSON: Great.

So, it is 10 after 11:00 now, and I think we have been going for a while. I don't know about you all, but I need a break.

(Laughter.)

So, why don't we reconvene at about 11:20 and we can go on?

(Whereupon, the foregoing matter went off the record at 11:07 a.m. and went back on the record at 11:22 a.m.)

CO-CHAIR PENSON: All right. So, not everyone is back in the room. But if start, I suspect people will come right back on in.

So, basically, we are now going to get into sort of the scientific acceptability piece of this. I think that this will be an opportunity to get into some of these issues that have been coming up around risk adjustment, around what is measured, specifically on the validity of the piece.

And this is, also, I know that NQF had a biostatistical review. And so, I know that the experts are in the room for that as well.

So, what we will do is we will start with the reliability piece, as you can see there, at the various subgroups. And then, we will go on from there.

So, we will start with criterion 2a1, which is the construction logic and also goes into the various other pieces about clinical logic and various adjustments.

And the folks who we have starting with that are Bill and David Redfearn.

So, Bill, if you want to start? And, David, you can chime-in as well.

MEMBER WEINTRAUB: Sure. Thank you.

We have actually spent a fair amount of time discussing the construction already. So, I was prepared to sort of review it, but I will go through it extremely briefly.

So, this is a measure of Part A and Part B Medicare payments, using it as a proxy for cost. It starts three days before the hospitalization, includes the hospitalization, and 30 days after the hospitalization. So, overall, it is sort of a logical construction.

I didn't understand the reason for exclusion of Part D until the developer told us. We can discuss that some more. I, too, was troubled by it until I heard the reason.

There are other exclusions as well, including deaths and transfers. And I would like to hear a little bit more from the developer about the logic for that as well.

And then, the rest is more down in the biostatistical weeds, which I am not going to go through.

Overall, the ratings by the reviewers were for construction logic. So, these are group high and moderate, 20; 1 low. Clinical logic, high/moderate, 20; zero low. Adjustments for comparability, 21 and zero. Adjustments for comparability, 21. Costing methods, 21 and zero. Adjustments for comparability 21 and zero.

Reliability, a little more concern here. High, 7; moderate, 11; 2 low, and 1 insufficient. Overall ratings, high, 7; 10

moderate; 4 low, and none for inadequate.

The first comment, I am like, hmmm, you might think I like this measure. In some respects, I do, but there are some issues.

And I think that a lot of the issues that people had have already been placed on the table. Here, how concurrent conditions, high-complexity diagnosis, and planned readmissions, and other events be managed. How are we going to handle attribution within 30 days? We have really already discussed that. Why the inclusion of those transitions to Medicaid within the episode? And it is sort of more details like that.

The statistical review obviously was very, very important here. The R-squared and calibration statistics are acceptable, although more information like the R-squared per strata would be useful. The model appears to have a small, but consistent bias, with

predicted values being lower than observed than from GSK. Actually, predicted values are higher than observed. So, there is a little bit of a bias.

Overall, this looked to be a measure which is well-validated, which I think that is overall true. I think David is going to have more concerns about that with risk adjustment. I will let him address that.

Here is essentially that comment: "Using a period of 90 days before hospitalization for risk adjustment isn't enough time to adequately the patient's comorbidities."

David, I suspect that was from you. Oh, okay, more coming. Okay.

MEMBER REDFEARN: I wasn't the only one.

MEMBER WEINTRAUB: All right. All right, there you go. I will let you address that in admission.

I don't understand why the

rationale for Medicare Advantage is excluded. We will let the developer comment with that.

There are several problems with the reliability analysis. The authors make an untrue statement regarding risk adjustment based on the distribution of spending and R-squared by a decile. I believe that the MSPB risk-adjustment methodology is robust and fits consistently across deciles. We can let the developer comment on that as well.

So, I think that those are some of the major comments. One comment here is: "Not enough robustness testing. Marginal reports by statistical consultant."

CO-CHAIR PENSON: Great.

MEMBER WEINTRAUB: Yes, I will stop there.

CO-CHAIR PENSON: So, David, do you want to add to that?

Again, it is hard for me, again, to sort of remember we are starting with reliability and then validity. So, in certain

respects, a lot of what we are talking about with the risk adjustment issue, that comes into the validity piece more than the reliability piece, but, of course, it spans both.

MEMBER REDFEARN: Yes, there are three comments here. My comment is the one at the very end of this, but there are three comments here, all pointing to the thing.

There are two components to the issue that I think we are talking about One is, is HCC the right model to use to do the adjustment? And I realize that I am kind of swimming upstream when I say I don't think it is a good model. It is the CMS model. It is what is used for this. So, that is probably not the most important concern here.

But there three comments in here, including mine, that using only 90 days of that member claim experience to do that calculation for the HCC is very likely inadequate. It is going to necessarily

underestimate the overall risk component for that patient. It is going to under-adjust for that component of this analysis, and that is the concern.

Perhaps -- I am making a guess here -- perhaps the intent was to only use 90 days, in the hope that you are going to capture the clinical conditions that are more closely related to the actual admission. And I can understand that that may be the case, but I don't think that is a very strong reason for restricting it to just 90 days. And that is just the kind of core concern, and several people I noticed had that same kind of concern.

CO-CHAIR PENSON: Go ahead, Brent.

MEMBER ASPLIN: Do you consider that -- kind of to the other David's points -- I agree and I had the same comment, but I put that more in the validity section than reliability. I am just wondering if you could comment specifically if you have any concerns

in the reliability aspect of those issues.

MEMBER REDFEARN: I think it is predominantly validity.

MEMBER ASPLIN: Right.

MEMBER REDFEARN: But I think it can affect reliability, too.

One of the ways that you can look at these risk models that have a defined period of time that is less than the full 12 months, you can say it is an opportunity. When you are scanning through a bunch of data looking for the diagnosis codes that really drive these models, if you look at a constrained period of time, you may not see the diagnosis codes that are actually present, simply because they haven't been coded in that period. And I think that can affect the kind of reliability, the end reliability of that score, because you just didn't happen to capture it.

It is there; it is existing, but perhaps the patient is not seeing the doctor;

they are only seeing the doctor very four months during the year. And you just never had an opportunity to see that diagnosis code. So, you are not going to generate the HCC. It is not going to go into the risk model, and perhaps it blends into reliability as well.

CO-CHAIR PENSON: Go ahead. Go ahead.

MR. AMIN: David, it might just be helpful -- sorry -- just to this point: specifically what we may want to focus on for reliability, because there is an overlap and a blend here, but we are only talking about the data element level. But let's try to keep this to whether the specifications are precise. And then, when we go into the conversations about validity, you get into a little more about the appropriateness of these various different components.

CO-CHAIR PENSON: Yes, I think that is very helpful.

David?

MEMBER GIFFORD: Just tell me if this is a later section. The preciseness of the definition of long-term care using the prior 90-day period, I didn't see it really clear in there, and I couldn't tell whether it is SNF Part A or long stay with a Part B in there. And those are different populations that actually would have a very different issue on the amount, the cost of a SNF stay versus someone who is a long stay being hospitalized. They are very different.

CO-CHAIR PENSON: Yes, I think that comes here. So, I will ask the folks from Acumen to comment on that.

MR. ZAIDI: So, the definition of long-term institutionalization is the same that CMS uses to risk-adjust Part C payments. And so, that is we use the MDS assessment. So, it is anybody who is in a long-term facility, either SNF or just a regular nursing home, for more than 90 days.

MEMBER GIFFORD: So, then,

individuals who had prior hospitalization, in per SNF and, then, went home would not be in that risk-adjusted group?

MR. ZAIDI: Not for this specific risk-adjustment variable.

CO-CHAIR PENSON: Nancy, you have your flag up there.

MEMBER GARRETT: Yes, I have a question, actually, for David. So, I also was concerned about that 90-day period not being adequate to capture enough of the diagnoses, but it looks like the measure developers actually did some modeling and they tried using a full year, and they didn't see a difference, much of a difference, in terms of how the results came out.

So, I wanted to get your reaction to that analysis. Does that change your mind at all about that? Or what do you think about the analysis stated?

MEMBER REDFEARN: It doesn't change my mind. I mean, I didn't see that

portion of the analysis, but everything that I have seen in terms of these models is that there is a really dramatic reduction in R-squareds when you drop below seven months. And we are only talking about three months here. That is a very dramatic subset.

MEMBER GARRETT: Yes.

MEMBER REDFEARN: And everything I have seen indicates to me that the predictive power is reduced dramatically.

MEMBER GARRETT: Right. So, can the measure developers comment on that?

MR. ZAIDI: Yes. So, we have seen the --

CO-CHAIR PENSON: Turn off your microphone there.

MR. ZAIDI: So, we are using the HCC model for a different purpose than what you might be used to. We are using it to predict short-term acute costs surrounding a hospitalization. And so, we actually did sensitivity analyses, as you mentioned, where

we looked at how would the R-squared change if we had longer look-backs. So, 120 days or a full year. And we actually found, to our surprise, that the R-squared was better with a 90-day look-back.

So, what that suggests to us is that the HCCs from the prior 90 days are more relevant to the cost drivers in the short-term acute-care setting than HCCs from a year ago.

But you might be right; when you are trying to predict annual cost, it is better to have more of a history. But, for this specific purpose of predicting the cost surrounding a hospitalization, we found that it was actually the model performed better with only 90 days of look-back and you get more patients, too, more eligible patients.

CO-CHAIR PENSON: So, I think that, just to summarize that, I think you are going to have to, the Committee is going to have to decide if that is acceptable with regard to reliability because you have heard

the measure developer's rationale.

Mary Ann? And then, Cheryl.

MEMBER CLARK: Yes, I just had a comment. It seems like there is a little bit of an inconsistency in, I know in the more detailed backup information, you are defining the full cost associated with all Medicare Part A and Part B services. But, yet, in the resource use categories defined in this application, it looks like some of them are missing. So, maybe that was just inadvertently they were left off. For example, home health and skilled nursing and hospice and a lot of different categories of cost.

MR. ZAIDI: Sorry. Where were they missing?

MEMBER CLARK: Under the resource use specifications.

MR. ZAIDI: Oh, that may have been an oversight then. It should be included.

MEMBER CLARK: Okay. And then, I

had a question about the timeframe used to capture the cost for the actual episode cost calculation. It looks like it was a certain defined time period of claims data being used from -- what? -- May through January, I believe, right? Is that going to be a consistent -- that is kind of the timeframe that is going to be used going forward? And if so, how do you reconcile with the standardized payment amounts?

I mean, there are two different fiscal years in effect there. So, are you taking that into account? Or is it one fiscal year for the pricing, the standardized pricing that you are using?

MR. ZAIDI: So, we are using the payment rates in effect at the time of the claim. So, if it is in the next fiscal year, it will have the payment rates for that fiscal year.

We don't think there is any bias. The only way there would be bias arising from

that is if some hospitals have a lot more of their admissions in one part of the performance period than others. We didn't test for that specifically, but, yes, the official CMS standardization methodology uses the payment rates that are in effect at the time of the claim. Yes, it changes by fiscal year.

CO-CHAIR PENSON: Cheryl?

MEMBER DAMBERG: I had two concerns with respect to reliability. So, when I was looking at the test/retest, I found it concerning that 30 percent of those that you would have scored in the high-cost, I guess it is quintile, would not be there the next time around. And that seems very high. I was wondering if you could comment on why there is so much shifting around.

And I am wondering to some extent, I mean, cost data is very noisy. My second concern, which I think relates to the first, is that in the information that you provided

around the reliability scoring you note that .4 is the lower limit of moderate reliability. And usually, that signals very poor reliability, and there is far more noise than signal in that estimate.

And I know you set the minimum threshold at 25, but I am wondering whether the reliability would be significantly strengthened if you bump it up to 50 cases.

MR. ZAIDI: Okay. So, I will take those in term.

So, the first on the quintile stability analysis, I think a lot of that is because they just, in the retest sample, they will be in the fourth quintile. So, it is like just below the cutoff for the fifth quintile.

So, if you look at the actual table we had, almost all hospitals that were in the top quintile in one sample were in the top two quintiles in the other sample. And the Spearman rank correlation between the two

samples was .8 or .835, which we thought was very high.

As for the reliability scores, 62 percent of hospitals had a reliability score over .9, and 98 percent had a reliability score of greater than .4. And the overall average reliability score was .95. So, we can debate what the minimum reliability score is, whether it was .4 or not.

MEMBER DAMBERG: Right, but I think in terms of the application, I guess if I were CMS, I would be hesitant to use or report anybody's information who had a reliability probably less than .8 on this particular measure.

CO-CHAIR PENSON: So, Bill, I know you have your hand up there. But, before I let you talk, I am going to ask Andrew, who reviewed the -- I am sorry to put you on the spot. You reviewed the reliability testing portion, and I wanted you to weigh-in, since we are sort of getting into the weeds there.

MEMBER RYAN: Great. So, there were a number of analyses that were performed related to reliability testing. The first that was noted was this test/retest reliability which was just discussed.

And the kind of headline number there is that 70 percent of hospitals in the highest quintile and 70 percent of hospitals in the lowest quintile of cost remained so through retesting. Actually, I thought this was quite a good idea. I like this analysis.

So, the other thing that Acumen did was this seasonality testing, which I think is kind of questionable. They said this is related to reliability, et al.

The one question I had about the reliability score was kind of how this was calculated at the level of the hospital because your reliability is typically thought of as at the level of the measure rather than the hospital.

But the total score was quite

high, over 95.95. But, again, as the authors note, this is really driven by the large number of average episodes across the hospitals. And so, the way the measure is calculated is that, basically, if there is any kind of between-hospital variation, and you have got a large "N", you know, you are going to have a high reliability. So, I don't get too excited about those numbers. I like test/retest better in this case.

And then, the bootstrapping analysis basically tried to get at this idea of, do changing cutoffs, how does that affect the confidence intervals for kind of the average hospital in the sample? But I think that here it would have also been good -- and I wonder if Acumen did any of this -- to see kind of, based on these confidence intervals that were constructed through bootstrapping, kind of what proportion of hospitals were statistically different from the kind of mean hospital in terms of their performance, to see

if there was any kind -- you know, there is variation, but can we identify kind of statistically-different variation? So, I think that would have been a nice addition.

Overall, I think I would have liked to have seen a little more with the testing. It was a couple of pages. It seems, when you look at the scores that the TAP gave, it was high on everything. And then, the reliability testing, it was more mixed, and the overall rating seemed to kind of line up well with the reliability testing. And I think some of us were a little disappointed with kind of how thin this was.

And then, you know, some other ideas I had which would be good to show is, you know, when you think about this episode, there is a couple of different components. So, there is the different costs that could be included in the episode, the kind of different risk-adjustment methods that have been discussed a little bit, and then, the kind of

length of episode pre- and post-hospitalization.

And then, I think it would have been nice to show some analysis of reliability across somewhat varying measure specifications to give a flavor as to why the 30-day interval post-discharge is giving us a more reliable signal than some other window, and why that 90-day, you know, apart from the R-squared analysis, why that 90-day look-back was the right interval to cut for the HCCs. You know, to show that in terms of other reliability analysis would have been good.

So, those are my summary comments.

MR. ZAIDI: Thank you. We really appreciate that feedback.

I think some of these comments were included in the statistician's review, and we had a chance to conduct some further analyses in response to the statistician's review. And that is included in the memorandum. I think it was one of the

handouts.

So, we did compare a 90-day post-discharge window with a 30-day post-discharge window, and the correlation was extremely high, over .85. So, we found that the length of the post-discharge window didn't have an effect on hospital ranking.

Another test that we did subsequent to the statistician's review is we did a year-on-year correlation. So, we didn't have two years of data when we first did the submission, but now we have two performance periods. And we did a correlation between hospital scores between the two periods, and we had a statistically-significant correlation of .8, which we thought was very positive and statistically-significant. That is the Pearson correlation.

MS. WILBON: I just want to note briefly that Carlos Alzola is actually here, who did the statistical assessments. So, if you have any specific questions for him --

sorry, we have been caught up in the mix, and we weren't able to introduce him earlier. So, if you have any specific questions for him about the assessment that you reviewed in terms of your interpretation of that, and how it plays into the discussion, he is available for that as well. So, I just wanted to make everyone aware of that.

CO-CHAIR PENSON: So, Bill, you have been patiently waiting.

MEMBER WEINTRAUB: I also have some questions for the developer. I can well imagine reasons for exclusions, death, transfers, Medicare Advantage, and so forth, but I would like to hear from you.

Also, unless I am missing something, cost is not generally normally distributed, and very often, in cost models the dependent variable is a log of costs rather than cost. So, if Carlos and the developer can comment on your choice of dependent variable?

MR. ZAIDI: We didn't test modeling log cost versus the levels of cost. That is an analysis we can do in the future. But I think the R-squareds that we have got with modeling the levels of cost were sufficiently high, but we can do that analysis in the future.

MEMBER WEINTRAUB: It would be worth it if it turns out to be better with log cost.

MR. ZAIDI: Definitely.

CO-CHAIR PENSON: Jack, before I let you go ahead, I am just going to ask Carlos to say a few words about the memorandum, and if you thought it addressed some of your concerns adequately or if you have remaining concerns.

MR. ALZOLA: Thank you.

I was just looking at this memo as I arrived here and trying to look at the answers that the developers gave to my issues, to the questions I raised.

To be quite honest, don't let me try to address all the issues that people have raised so far. I don't have any issues with the reliability. I know that you may think that reliability scores in some cases may be low, but, by and large, they are quite significant -- they seem to me pretty large.

And besides, their reference is quoted here, that this is considered, in all cases, the reliability is about the level of what is considered moderate reliability, and there is a reference provided for that.

The other issue, the thing that really makes me satisfied about the reliability testing is the test/retest reliability. I basically look at the Spearman correlation, and .85 to me is more than good enough.

So, that means that -- maybe I should explain what that means. The ranking of the hospitals in terms of their MSPB in the sample and in the retest sample, which is more

or less equivalent because it is a 50/50 random split, it is pretty much the same. Of course, it is never going to be equal, but some hospitals will run lower; some will run higher. But, by and large, they have the same ranks. So, that is why the high Spearman correlation makes me happy about reliability.

The second issue that was mentioned, that was raised often, was the testing, the look-back period of 90 days versus one year for the HCCs. I am of the mind that the closer you get to your hospitalization, the more likely those conditions are going to be to influence the care that you get. So, I thought that 90 days was sufficient, but you also have somebody who questioned whether that was valid or not.

And the testing that the developer performed about that, which was using the 90-day period and using the one-year period, gave basically the same R-squared. So, I am pretty sure that, if we were going to test the

difference of one R-squared versus the other one, it may be significant because of the large number of cases, but it is clinically totally irrelevant.

So, the big question -- and I think I am going to need to ask the developers about this -- was about validity.

CO-CHAIR PENSON: So, why don't we --

MR. ALZOLA: Shall we wait for that?

CO-CHAIR PENSON: Yes, why don't we wait on that?

MR. ALZOLA: Yes.

CO-CHAIR PENSON: So, Jack, you have --

MEMBER NEEDLEMAN: Jennifer and a couple of people were in the queue ahead of me. So, I don't want to jump the queue.

CO-CHAIR PENSON: Okay. No worries. Jennifer go ahead.

I am not allowed to ask your

question. Okay?

(Laughter.)

MEMBER EAMES-HUFF: So, I guess I have a couple of questions here regarding the exclusions that you have used in the measure. And I am coming from the perspective of I am wondering if we potentially are unnecessarily excluding patients with high expenditures, which may skew the results of the hospitals, but also lose some opportunities for improvement.

And the two exclusions I would like to discuss more is the one around transfers and deaths. I think it was the transfers that count for about 5 percent of the population, which seemed like some level of significance in this population, and that it wasn't clear how to attribute the transfers. And so, that is the reason why they are being excluded. I wanted to see if you have further plans for resolving the attribution or if this is going to be an

exclusion sort of indefinitely, and whether or not you considered if it is really the role of the different hospitals, and if there was any consideration around attributing to both hospitals and having some level of shared responsibility and accountability for this.

And in terms of the death, if I am understanding it correctly, it is essentially saying, because the patient wasn't around for the entire episode, you are not capturing all their costs. But it feels like they are around for as much as they can be, and end-of-life is really important area. So, I didn't understand why patients that died were removed.

MEMBER WALKER: Jennifer, I have a question for you. I raised those exact same questions, but in the validity section. So, is this a question about the reliability of the measure or about the validity of the measure?

MEMBER EAMES-HUFF: Yes, I was

following the 2a1 that says comparability for inclusion/exclusion criteria. So, that is why I brought it up in this section.

CO-CHAIR PENSON: Yes, and you have the same structure for the validity, too. I think, with this, you really want to look at reliability vis-a-vis capturing the mechanics.

MEMBER EAMES-HUFF: Okay. So, we can do it in validity.

CO-CHAIR PENSON: So, Jack? And then, Bill.

MEMBER EAMES-HUFF: It doesn't matter to me as long as you get to it at some point.

CO-CHAIR PENSON: I think we will get to it in the validity. I think if we sort of take the reliability sort of as the mechanics and the math, and then, take the validity as the sort of underlying constructs, and does it flaw the results, is a good way to look at it.

Jack?

MEMBER NEEDLEMAN: Yes, I share Cheryl's concern about that 70 percent, which would say 30 percent of the folks in the top quintile are different. So, I think that may be a validity issue.

But, in terms of the construction, this winds up on the Hospital Compare site with hospitals better than average, average, worse than average. So, how does that 20 percent -- where is the cutoff for better than average in this measure? Is it at the quintile level? Is it at the decile level? Because we ought to be looking at reliability in terms of consistent classification around the point at which people are going to be put into the worst category or the average category.

CO-CHAIR PENSON: So, I think that goes to either CMS or the developer.

MS. SPALDING BUSH: So, currently, on Hospital Compare, the measure is just displayed as a ratio to the national median

and it is not bucketed into better, worse, or average. It is rounded, but there is no cutoff where we bucket them.

There is a statement there that says something like: a ratio that is greater than 1 means Medicare spends more per patient at this hospital than the national average, but they are not put into buckets. That is just to help consumers understand that a higher number means more Medicare spending per episode.

MR. ZAIDI: Can I make one comment? For hospital value-based purchasing, the measure is going to be evaluated on a continuous basis. So, it is not like there are attainment points and improvement points that are continuous in nature. So, there aren't these kind of quartile buckets that hospital will be put on in these measures.

CO-CHAIR PENSON: So, here is what I would propose: I think we can call the question on this with regard to reliability.

And I want to remind people that we are really voting on the preciseness and the ability to put things -- you know, the precision of constructing the model, whether their test/retest works. And the actual validity and does it measure the constructs, we will get to in the next section, if everyone is okay with that.

So, unless there are other comments, I would suggest we take an overall vote on reliability.

MS. WILBON: So, before we vote, I just want to make sure that people are clear. Reliability includes reliability of the specification. So, to what degree you believe the specifications are repeatable and can be implemented consistently, as well as the reliability testing. So, we are going to vote on overall reliability for the measure, which includes both of those components. So, I just want to make sure that everyone is clear on that.

And within the testing section of that, of the reliability criteria, we are looking at the appropriate method was used to test the measure; the scope of the testing was adequate, and the results were within acceptable norms.

So, if everyone feels that those items have been discussed within the testing and that you have a sense of how precise and repeatable the specifications are, then we should go ahead and vote, keeping those two things in mind.

CO-CHAIR PENSON: So, on the basis of that, any additional discussion?

(No response.)

All right. Hearing none, let's have a vote then.

MR. WILLIAMSON: We will now vote on the overall reliability. You will have 60 seconds. Please begin now.

(Vote taken.)

We have 10 high, 14 moderate, 1

low, and zero insufficient.

CO-CHAIR PENSON: Great. So, I was just talking to Ashlie. What I think we will do is -- we are pretty far behind on time, not that it is the end of the world because tomorrow is a fairly loose schedule. But, that being said, what I would propose we do is turn this into an -- I may get food -- but a bit of a working lunch. So, maybe take 15 minutes to get your lunch, come back, get started eating, and we can go through the next section, which I think will be some work.

(Whereupon, the foregoing matter went off the record at 11:59 a.m. and went back on the record at 12:27 p.m. for a working lunch.)

A-F-T-E-R-N-O-O-N S-E-S-S-I-O-N

12:27 p.m.

DR. PENSON: So let's get started again because I think that this next section on validity will be an area where we have a lot of discussion. And a lot of the issues which were discussed on the reliability section really are going to come up again here. And this is where they should come up.

So as you can see in front of you there are a whole lot of sort of subcategories that we're going to go through. And what I think we'll do is we'll go through them each one at a time, let people talk.

And just remember as you can see we have issues, validity of construction logic, clinical logic, adjustments for comparability with inclusion/exclusion criteria, exclusions, adjustments for comparability with regard to risk adjustment. Then the actual risk adjustment piece which -- and costing method, scoring method, et cetera.

So there's going to be a lot of ground to cover and a lot of folks have gone through the different parts so hopefully we'll be able to do this in a way that we can really do it systematically and come to a good vote.

And also at the end before we vote as Gene pointed out and he'll remind me but others should as well we'll invite public comment from people in the room or on the phone who are interested which I think this part will be important.

So why don't we start talking with 2b1 which is the validity of the specifications. And we're looking at things that basically are on are the specifications consistent with the measure intent with regard to construction logic, clinical logic, the various adjustments for comparability and the measure score calculation.

So we did assign these out. And so let's start by talking about sort of the validity of the specifications. And David and

Martin reviewed that. So I don't know if one or both of you want to get us started.

DR. MARCINIAK: Sure. Why don't I take the first whack at it and then we'll turn it over to David.

Sort of at a high level, you know, David has pointed out that this is the section about measure specification. Are they consistent with the measure's intent described under criterion 1c the intent of the resource use measure and the measure's construct, are they clearly described and does it capture the most inclusive target population. So what's embedded in here are really sort of the three core modules. It's the construction logic, it's the clinical logic and it's the adjustment for comparability.

I know that we've had a large conversation around parts of this this morning. We were starting to bleed into it so I'm not going to talk to a great extent about a lot of the details at this point, assuming

that they'll come out in the dialogue.

When you look at the range of voting, so for construction logic probably skewed to the left in the high to moderate area, that we had 19 votes there for construction logic. Two were voting low.

When we looked at clinical logic again skewed to the high to moderate. Nineteen versus two for low.

And then when we looked at adjustments for comparability, particularly around inclusion/exclusion we again saw that same sort of framework in the vote. So it's a leftward skew to high/moderate.

Overall there seemed to be good consensus. The clinical logic and the adjustments with respect to clinical logic and adjustments for comparability. The main concern appeared to be really on risk adjustment and validity testing for both which will be covered probably a bit later in our discussion this afternoon.

So, I'm going to turn it over to David to fill in any egregious gaps in my conversation before we open it for discussion I guess.

DR. GIFFORD: I really don't have much to add to what Martin has said other than I think they did a nice job laying out the logic they had there.

I think some of the issues in the logic really relate better to the discussion about the validity or the exclusions in the validity and later on. And some of the issues in the construction logic we've already talked about earlier on and I don't want to beat that dead horse any further as we go forward on that. But it was well laid out and made sense.

The only -- I'll kick the dead horse quickly and just say they continue to repeat that this is about care coordination and I think that that doesn't come through in the logic they show and argue there. And that

would be the only big piece I'd make.

DR. PENSON: Well, if nothing else I'm feeling better that I didn't make that up, so that's good.

So I think this is a good time perhaps to open up a general discussion. We're going to talk a little bit about the actual statistical testing in a minute, just the way things are laid out.

You know, it might be valuable when I look at the list here that there's people, there was some concerns. I mean if you look at the way people talked about exclusions. And you see that a lot of people beforehand said they were moderate. There wasn't a lot of high there. So I wonder, Cheryl, you had -- pardon me, Herbert, you had talked a little bit about the exclusion so I wonder if you want to share your thoughts on that.

DR. WONG: Yes. So I also recognize if you take a look at the

distribution with the exclusion we had high -- five of the folks voted high. Most folks voted it moderate and then you have two folks that are basically on the tail end.

And to a large extent I think that most of the comments were actually quite silent on the matter. However, we talked a little bit about the exclusions a little bit earlier on. And I think that from that perspective it -- those are the key comments that kind of percolated up.

One comment was just a step back, thinking about this measure in general. Should it be an overall resource measure and not specific to Medicare in itself. And the comment about exclusion was this notion should one also include other payers that are part of this episode in some way.

So for instance, what about transfers? How do you deal with that? How do you deal with death? How do you deal with Medicaid payments once Medicare expenditures

have basically expired? So I think that those comments were kind of percolated up a little bit earlier.

I think that there are another notion about exclusion. And I think that this spills over a little bit more on the validity testing. But it was kind of briefly mentioned in the exclusion commentary.

One of the notions was whether or not that when you take a look at the validity testing the methods that were employed are basically designed to kind of look at a broader aggregate analysis.

And the question emerged of whether or not this is well designed for individual analysis. And so as part of that you do have certain cutoffs, certain exclusions. And the person who made this particular comment kind of wondered whether or not these cutoff thresholds would actually have an impact when the measures are designed for a different purpose as opposed to

employing certain methodological approaches that are designed for a different reason.

Somewhat related to this was this notion about when -- how were the exclusions kind of performed. And someone recognized that there were roughly 89,000 outliers. And that called for some concern of whether or not the models themselves were in fact valid.

And to the extent that folks did not have the sensitivity analysis and things of that nature where the exclusion really designed to kind of make it more -- make it look more appealing than others. So I think that that's how -- my interpretation of the commentary there.

DR. PENSON: So I'm hearing some concerns about the exclusions although I'm not hearing that they're complete deal-breakers. A number of times you alluded to the testing. And so I think it may be a good time to have Cheryl comment on the validity testing and maybe also Carlos as well. If you want to

comment on validity testing.

DR. DAMBERG: In terms of what's described under 2b2 which was the area that I was supposed to address they looked at two tests of validity. One was looking at the correlation or association between the measure and the percent of beneficiaries with multiple episodes.

And the second test was to correlate the measure with other outcome measures, namely hospital readmission measures. And what they found was generally very weak correlation with both of those measures. And in particular, the hospital readmission measures, the correlation coefficients were 0.8. I'm sorry -- 0.08, 0.07 and 0.06.

And although there were comments that they gave in response to some of our comments so I should highlight those as well. They say, "We wish to clarify that there is a weak positive but statistically significant

correlation with each of the readmission measures." So that was sort of an add-on to what was included in our documentation.

I think in terms of how the group scored this for validity testing high was only 1, moderate was 11, low was 7 and insufficient was 2. And the issues that I would call out from the group, there were about five different issues.

I think pretty much across the board folks commented that these were not particularly good tests of the construct validity and suggested some other approaches that could be used.

The second issue that came up, and this has come up throughout the day was this issue around exclusion criteria. And wanting to better understand why we were excluding transfers, deaths, getting back to Jennifer's issues, as well as Medicaid payments. So that was the second concern.

The third was around the risk

adjustment model validation. And I think we've heard a fair amount about that today as well.

And the fourth issue was raised around what they considered to be an unstable model. And concern about possible co-linearity among some of the variables in the model.

And then another issue that was flagged because of the lack of adjustment for age and race is the potential for unintended consequences. Would this particular measure incentivize hospitals to avoid high-risk patients basically that they may have a harder time controlling the costs on. So I think generally a number of concerns and fairly low ratings for this particular measure.

DR. PENSON: So I wonder if other folks have thoughts about the testing and the exclusions. Go ahead, Brent.

DR. ASPLIN: Well, I would like to hear -- I can understand in some respects why

in this particular version of the measure some of these things were excluded, particularly the transfers and deaths. How are we going to manage that?

I mean because if you think about perverse incentives and trying to understand probably our highest cost acute care episodes, those are it. Transfers and deaths are going to be the most costly acute care episodes outside of some very high-end planned transplant or some other things like that. That's where the money is.

And in particular if you're a community hospital that is on the hook for this over time and you see a high-cost patient who isn't doing all that well, why not transfer him to Jim's shop in Rochester, right? If you're in southeast Minnesota. Other than the fact that all the hospitals you refer are part of Mayo too.

(Laughter)

DR. ASPLIN: You know, why not

unload them on a referring center? So it kind of comes back to some of the usability and maybe other issues that get separated. I don't think it's a deal-breaker yet it would be unfortunate if we don't try to grasp and grapple with those issues over time. Because they're very important.

DR. PENSON: So let me turn that question around because I think it's a very valid comment. To summarize what I think you're saying is that because of the exclusion criteria there's the potential for gaming the measure both consciously and perhaps subconsciously.

So I'll ask either the folks from CMS or Acumen to share their thoughts about that.

MS. SPALDING BUSH: So, we appreciate the comment and it's something that we did consider. When we first put the measure forth through rulemaking we did not have an exclusion for transfers.

We received a number of public comments on this particular issue on both sides. So basically you heard community hospitals saying what are you going to do about transfers. You can't have me on the hook for this patient that I just maybe stabilized but knew was, you know, outside of my realm of expertise and so I transferred it to another medical center.

We had comments from the receiving medical centers sort of saying, well, there may have been mistakes made at the community hospital and that secondary to that that those patients may have lived closer to that community hospital. So their follow-up care is going to be outside of the receiving hospital's sort of realm of influence.

So I think that while the potential does exist for -- because we've excluded these that there could be an incentive for the admitting hospital to transfer a patient I think that same incentive

would exist were we to attribute the episode to the receiving hospital. So I mean the only other option there would be to attribute it to the community hospital and then have them on the hook for this very complex patient that they potentially just didn't have the ability to care for. And so that's why they wound up excluded.

And I think someone's commented here that this was the least worst approach or something like that. I mean we wouldn't have used those words but I think it does come down to that. It is sort of what's the best thing that we could do in this measure.

And I think the other issue with attributing it to the community hospital is would we give them pause, then, to not transfer a complex patient to see if they could manage them rather than send them away because they were afraid that they were going to go to an expensive medical center and incur some huge cost that that community hospital

didn't want to be held accountable for.

This is something that we can consider and we plan to continue evaluating for maybe a future refinement of the measure. But I think for the initial implementation it seemed like the best solution given the public comment that I think was really valid that we received on the measure.

DR. NELSON: As a point of clarification can you tell us about what proportion of episodes involved transfers and what proportion of dollars involved transfers?

MR. ZAIDI: It's about 5 percent of hospitalizations overall involve a transfer. And I'll have to look up the dollar amount.

DR. PENSON: So I see Andrea has her hand up and then Jack and Andrew. Okay, Andrew, do you want to go first? I hear you were first. We'll get to everyone, I promise.

DR. RYAN: I think the point that was just made about gaming the measures is

great. And you could imagine hospitals kind of transferring tough patients back and forth between each other to both get off the hook. I mean I'm kind of -- it's hypothetically possible. And so if the stakes were high enough hospitals could do that.

The logic that was just given that why should the community hospital bear responsibility after the transfer isn't persuasive to me because we're asking them to bear responsibility for 30 days after, you know, everything that happens 30 days after the discharge. So it seems like they do control that transfer. And the idea of trying to make more high-value transfers seems reasonable to ask hospitals to do if we're thinking about making them accountable for everything that happens in the 30 days.

And then two other points I wanted to make. The end of life spending and excluding deaths seems like a real serious issue because you can imagine a high-intensity

hospital that keeps patients alive could look worse on this measure than a high-intensity hospital where patients actually die because those patients get excluded.

And then I didn't see any evidence that the lowest 1 percent of episodes were actually outliers. I imagine that 1 percent is probably pretty close to 2 percent in terms of the spending. And I just kind of -- I can clearly see that the top 1 percent is likely high-cost outliers. But I really do wonder how much it makes sense to exclude that bottom 1 percent. So that's it.

DR. PENSON: So just before Jack and then Andrea, I just, I think in the end when we talk about the transfer exclusion and to some degree the issue is death as well. I think in the end we have to take it at face value and you have to make your own judgment when you vote. Because I think that the comment that, you know, it could go either way is probably well taken.

So you have to decide if the way they went is what you like or not, if that makes some sense. So I think if the order is correct then -- oh Daniel, it was your question. Okay, so go ahead.

MR. WOLFSON: It's deja vu all over again. I mean when we were in the eighties we had utilization information. And we said well, there's unintended consequences. To beat that system just do less and you'll be fine. And that's why we developed quality measures. So it just seems like we've come full circle.

And you can't do these measures. We're just proving that you can't do these measures without quality measures because there's unintended consequences to game the system. So I think that we could go around and around but ultimately it goes back to the notion that you can't have these measures without quality measures alongside of it. We're all in agreement of that. So I think we

can go on with the unintended consequences.

We know there's unintended consequences, that's why we developed quality measures. That's why NQF is here and NCQA, to combat utilization-focused measures. A little history.

DR. PENSON: So let me just add to that and any of the NQF folks can chime in. Because Martin came to me during the break and made a point. For those of you who remember phase I of this project you'll remember the heated discussion that we had about this.

And the concept particularly on the provider side that you can't -- exactly what Daniel just said. And I actually made that very pressing argument except my comment was I don't trust anyone here. You're still going to end up making it all about utilization and you have to make it quality.

Unfortunately I think we're going to have to trust them. So otherwise we're not going to be able to get done here.

I understand people are uncomfortable with that. I'm uncomfortable with it as well at times. I'm going to guess that CMS is an honest broker to start with and go from there. So I just wanted to add that. I hear you, believe me, I hear you. Jack?

DR. NEEDLEMAN: Yes. I have a lot of questions, issues. I'm not quite sure what they are, which bucket they fall into.

One is this issue of the exclusion of deaths and transfers. Given the DRG payment, given the standardization of payment it's not the hospital portion of this that's at issue, it's the gaming by transferring patients that you think are going to need lots of post-acute care.

And we normally expect transfers to be based upon the ability of the hospital to deliver the acute care services that patients need. And certainly this does add a gaming component to that, but we need to think about how serious it is that hospitals are

going to transfer based upon assessments of the post-acute needs of these patients. So that's a question.

The deaths, we've got deaths excluded which is an important quality metric here as Daniel noted. But we've got hospice costs included. So I'm just wondering how consistent we're being here in what we're measuring and which patients we're looking at, which patients we're excluding.

So I would raise that issue in terms of the appropriate scope if deaths are included. And whether the hospice patients ought to be excluded as well. So that's a question I have about whether the exclusions are right.

As I think about the construction of this measure with the inclusion of an indicator variable for the DRG that the patient's in we've got something that in any of these regressions for the risk adjustment is going to pick up in essence the average DRG

payment for the patient's diagnosis. So what the measure is really measuring is variation between no post-acute services and lots of post-acute services because we've got the core cost associated or the expenditure associated with the hospitalization already included through the DRG in the risk adjuster.

So we're looking at variations in pre- and post-hospitalization acute care, expenses as the major sources of variation in this measure after risk adjustment. And the question is whether the HCC risk adjuster and the age adjuster and the other adjusters here are capturing everything that produces variation there. So that's one issue in terms of the risk adjustment in general.

But the more specific issue which I think comes up as we look at some of this data is whether the high-cost places are relatively consistent in which sets of post-acute services are driving you into those high -- being that high-cost group. And the high

R squared in that group suggests there's a lot of consistency in what's happening in that group.

And again, it's an issue of in terms of the construction has there been -- earlier we had talked about -- Acumen and CMS had pointed out yes, we see readmissions, we see skilled nursing facility, we see the outlier patients in the high-risk end of this. I know the correlation with readmissions is low but I'd like to know what the actual percentages are, particularly at the high-cost end of this measure where people are going to be very concerned and the value-based payment system is going to get very concerned.

How much of being a high-cost provider at the high end of the spectrum, not the full spectrum but the high end spectrum is associated with simply having a very high readmission rate or a very high outlier in terms of the patient inpatient experience rates. And whether the risk adjuster is fully

adjusting for which patients we can predict are going to show up in those categories.

DR. PENSON: So I think you've raised two points there. The first reiterates a comment you made earlier about what's driving the variation.

The other point that you've started a discussion is on risk adjustment which I think is an important case here. And I would also invite Nancy who was also looking at risk adjustment just to add your thoughts to Jack because we're moving along.

And I'll get everyone else's questions too, I promise.

DR. GARRETT: So it might be useful just to take a step back and get us all on the same page with what the risk adjustment methodology is I think, and then we can go into the issues. So if you wouldn't mind putting my slides up and go to slide 8.

So as we have already touched on the explanatory variables are looking 90 days

before the start of the episode. And so that's a point that we need to talk about some more I think. And so it includes severity of illness, some of the explanatory variables, severity of illness using 70 HCC indicators, an indicator of recent long-term care, disability or ESRD status. There's 12 categorical age variables and then also the DRG of the index admission.

And then there's an OLS regression model where standardized episode cost is the dependent variable. There's a separate model for each major diagnostic category that the DRGs roll up to. And so then there's an exclusion for the outliers at the 99th -- above 99th and below first percentile. And there's not an adjustment for race or for sex. So next slide.

So some of the key issues in terms of the comments. These are how the votes came out. One of them is the 90-day period, is that too short. The HCC model itself, that

it's weak compared to commercial groupers. And then not adjusting for socioeconomic factors, sex and race. And then there's some concerns about multicolinearity in terms of the correlations of the explanatory variables.

And that's -- those are kind of the key issues that I captured from the discussion.

DR. PENSON: So we've had a lot of discussion about risk adjustment. So I sort of am expecting comments. Cheryl and Jennifer, you have your --

DR. NEEDLEMAN: Can I just, as the other person on risk adjustment --

DR. PENSON: Oh, I'm sorry.

DR. NEEDLEMAN: -- can I just add one thing which I didn't do as well as I would have liked to do a few minutes ago when I was sort of quickly summarizing my reaction to the measure.

Which is if you think about the expenditure side of this it is the hospital

DRG payment, any outlier payment and everything else in Part A and B. If you think about the risk adjustment model which is being regressed on that it is the DRG and these HCCs and age categories and the other stuff.

Well, the core DRG payment is completely determined by the DRG category. So any variation you get in this measure that you're risk-adjusting for is the risk adjustment for the outlier payments and all the non-DRG payments. And everything which is sopping up that variance is the HCCs, the age and these other things. So that's what's going on.

Fundamentally if we're trying to explain variance through the risk-adjuster after the core DRG we're explaining that with these other -- with the other variables that are in here. And we've got to decide whether that makes us happy or less happy.

DR. PENSON: I think it's a key point. I mean I have to say that Nancy's

comments earlier about disparities resonated with me, that it just struck me that there were so many things that weren't accounted for in the risk adjustment that frankly wasn't completely fair. That certain hospitals serve populations and you know. I'm not a genius when it comes to HCCs, I haven't used them, but I get the impression from people that they don't capture everything in that domain.

Cheryl and then Jennifer or other way around.

DR. DAMBERG: Can I still talk about transfers? I just wanted to make one last comment on that.

So there's so many different market dynamics in effect. And this landscape is changing dramatically. And one of the things that I've observed, we're doing all these studies of nascent ACOs. And one of the things that's going on is hospitals that are part of these ACOs are reaching out and transferring patients back into their

institution to be able to manage those costs more tightly. And so they're repatriating patients.

And so I think that transfers shouldn't be excluded. And I think CMS really needs to monitor this. And I think if you start to see very strange changes in the readmission rates you need to look at whether that's an unintended consequence or whether it's due to some of these other market dynamics that are in play.

DR. PENSON: So I just want to make a point. I think that's a good point and it's interesting. I hadn't even considered that.

One of the things we have to remember as a committee is that -- or as a panel I should say is we have to sort of judge things as they are. So while I think it's helpful to give CMS advice on perhaps how to fix it in the end the vote really has to focus on what's been put in front of us. I just

want to remind folks of that. Jennifer?

MS. EAMES-HUFF: I'm also going to beat another issue to death but I have a different conclusion that hasn't been discussed.

So I will say I agree with everybody around the general premises that we need to report both cost and resource use measures in conjunction with quality. And that's really important.

I think I drew a different conclusion around the issue of since this doesn't have a quality measure given the history lesson of 30 years ago of looking at cost and resource and not having quality measures and then doing a focus on quality. That led me to believe if we don't move this forward with a quality measure do we have to wait another 30 years to get what we want. And it seems like moving this measure forward even without the quality piece actually will get the quality measure faster.

DR. PENSON: Great, thanks. So David and then Brent.

DR. REDFEARN: Just a quick question for the developers. Why was gender excluded? Normally you run these patient risk models, you have age, gender and then the clinical categories. Why was sex excluded?

MR. ZAIDI: Because as we understood it NQF policy on risk adjustment encourages not adjusting for --

DR. NEEDLEMAN: Yes, but you included it in 2165. So you've not been consistent.

MS. SPALDING BUSH: I think that could probably be better explained by that measure developer. But there were some differences in the way that our statutory requirements were written to start with. And we were not required by statute to adjust for sex so we went along with the NQF position and didn't.

MR. AMIN: So the NQF position on

sort of risk adjustment is specifically on where there's areas identified related to disparities. So that variables that are identified that are known disparities in the field should not be included in the risk adjustment model for obvious reasons, the fact that they would be masked in terms of the outcome.

So examples could include gender, well in this case SES in particular. So it's not to say that gender should always be excluded from the risk adjustment model. It's only when there's a known disparity between genders that you wouldn't want to mask that in a model.

So it's up to the committee in this case if that's appropriate as it always is. Even the issue of SES, it's still a question of appropriateness in the application.

DR. PENSON: Brent.

DR. ASPLIN: So we spoke quite a

bit about the transfer issue. I'll leave that one. But could the developer just speak to deaths? I mean -- and why they're excluded.

Now to the extent they're inpatient deaths and it's all going to be captured, the majority of the costs are captured by the MS-DRG there might not be much variability left over. You can certainly look at other scenarios where the death occurs within 30 days after discharge.

So I'm struggling with the rationale on deaths. I understand there's no good answer on the transfer issue. Help me.

MR. ZAIDI: So when we looked at the issue of excluding deaths what we were finding was it's kind of a bimodal distribution. On average episodes with death cost 40 percent more than the average episode.

But there were a lot of episodes where they were costing far under what you would otherwise predict. And we suspected that part of was because the patient died

early in the episode and didn't have the opportunity to experience the claims that other patients would experience. And because we thought there was no good way to adjust for that or account for that that was the reason we excluded deaths.

DR. PENSON: It strikes me as ironic in the discussion of the quality and utilization that this is a good example where you've got to get the quality piece. Because in this setting if your outcome was alive or dead it wouldn't work out well for you if you were the hospital.

DR. ASPLIN: That appears clinically totally what I would expect, a bimodal distribution. You know, you're going to have a group that clusters right early on, they're going to die right away and you're going to have a group that's going to have prolonged stays. That to me -- it doesn't concern me. It doesn't feel like a reason to exclude them.

DR. PENSON: So I think what I would say again with both the transfer and the death piece is that each of you is going to have to make up your own sort of judgment as to whether or not it's kosher for lack of a better way to put it. I am hearing a lot of discussion around that and so I think that'll come out in the wash.

Before we move along you can see we've talked about exclusions and a lot of these discussion points. And also with risk adjustment.

I wanted to just give Carlos a chance to say something. I wanted to bring this up before. You know, in your original review you made a comment about the validity testing being weak. And in the response document basically noted that there's a weak positive but statistically significant correlation.

And I wanted to get your thoughts on whether or not you thought the responses

were adequate and if it's changed your opinion, Carlos.

MR. ALZOLA: Okay. With respect to the correlation with readmissions for the three conditions, heart attack, pneumonia and heart failure I think yes. You argue that the correlations were statistically significant.

But in this context with a large, very, very large number of cases a statistically significant correlation is very, is going to happen anyway. What really matters here is whether those correlations are clinically significant. And I think that the values are pretty low.

You do argue here, and I thought that was a good point -- let me see if I can find it -- that the cost for the readmissions only included inpatient costs. So that goes to explain the local relation to some extent.

I still feel that there are other ways of testing validity that are more indicative of face validity. And I think you

do -- you did some work in that respect. And one of them was you calculated the correlations between two different periods which -- that's a step that I consider indicative of face validity. Because you don't expect things to change too much from one year to the next. So if things correlated in 2010 they should correlate over 2011. So it's stable over time.

And the other one that's -- the other step that you took was to correlate the MSPB measure with service counts of various procedural categories like procedure services, emergency services, and a number of other measures of resource use.

And you said that you found statistically significant and strong positive correlation with professional E&M services and post-acute services including inpatient I guess and SNF. The correlation was 0.6 which that's acceptable.

So I do -- I think that in terms

of validity those things help to increase my level of confidence. But I still would like -- personally I would still like to see more detail on those analyses.

DR. PENSON: So just to sort of reflect back because I think it's important. I think at least for me and I suspect for others in the room that that review really influenced where I came down.

What I hear you saying is that perhaps it's stronger than you originally suggested in your summary but you still have some concerns.

MR. ALZOLA: That's correct, yes.

DR. PENSON: All right, terrific. That's helpful to me. Why don't we keep moving down the lists. I think we've talked a lot about risk adjustment. I see these up. Nancy, David, do you have comments? Okay, go ahead. Sorry.

DR. GIFFORD: My comment is on risk adjustment. I'm assuming the long stay

in the risk adjustment is defined the same way.

About two-thirds to three-quarters of those will be dual eligibles. So I'm not clear. We're excluding dual eligibles. Then we're adjusting for long stay and whether they're in there or not. Because many long stay, two-thirds to three-quarters will have Medicaid and about 80 percent of them will be women. So in essence those measures are adjusting for the very things we're talking about. I was wondering why.

I mean, I agree with that should be in there but then it's not sort of consistent with everything else that's going on.

MR. ZAIDI: So yes, we definitely found in the data that being in a long-term care institution had a very strong effect on your episode cost. And there are always going to be -- a lot of different variables are going to be co-linear with each other. So

yes, we couldn't find a good way to separate out the effect of Medicaid from the effect of being in a long-term care institution. But because it is included in the CMS HCC model we thought it was appropriate to include that as a risk adjuster.

DR. GIFFORD: I guess I would agree but then following that logic with all of the other discussion we're having why aren't we adding other stuff in that we need to risk-adjust for? I guess that's my point I wanted to make. Either we have to be internally consistent or not. To pick and choose what we want to put in doesn't make sense.

DR. PENSON: I think, you know, I keep hearing concerns about risk adjustment now. The sort of random choice of what got in and what got out adds to it. Nancy?

DR. GARRETT: So I just wanted to bring up the point that we talked about this morning around the fact that episodes for dual

eligible beneficiaries were more costly but at the hospital level the results didn't change when they tried adjusting for that.

And to me that's a real validity concern. I don't understand that, what's going on. Is it some kind of -- I just feel like without knowing more about that I do have concerns about the validity of the measure. So now that we're in the validity section I wanted to bring that up again.

DR. PENSON: Thanks. Jack?

DR. NEEDLEMAN: I have a question for the developers. The more I keep turning this over the less -- I started out very happy with the risk adjustment. And the more I keep turning this over the more concerns I have.

And as I think about the decile results, as I look at that 0.46 R squared and the 0.6 R squared I realize you've got the core DRG in there which means you've got the core DRG payment which is a substantial portion of any of these expenses.

And I'm just wondering have you subtracted the standardized DRG payment from your total standardized payment level and regressed everything except the DRG? Regressed that residual on everything except the DRG to see how much R squared -- how much variance in the rest of the payments you're actually able to explain with the rest of the risk adjuster.

MR. ZAIDI: No, we didn't subtract the core DRG payment. And part of the reason is that we also include the CMS outlier payment which -- so even in the same DRG some cases will be slightly more expensive because they have that outlier payment. So it's not always exactly the same across all cases.

DR. NEEDLEMAN: Basically the variance around the core DRG payment and how much of that is in fact explained or can be risk-adjusted away. I suspect a lot of this 0.46 and this 0.6 is that core DRG payment.

MR. ZAIDI: So we haven't analyzed

the R squared excluding the core DRG payment but we'd be happy to do that analysis.

DR. PENSON: So Nancy, do you have another comment or no? Okay. So I'm hoping for the sake of time that we're sort of getting towards the end of the risk adjustment and exclusion piece.

I want to go onto 2b5 which Tom took a look at which was the identification of statistically significant differences. Tom, do you want to comment on that?

DR. TSANG: Yes. So I was asked to look at whether the data analysis demonstrates that methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically and clinically meaningful differences in performance.

And I think the key words here are really practically and clinically meaningful to me at least. You know, so on the preliminary scores that looked at this scored

with 5 in the high range and 10 in the moderate range and 3 in the low and 3 insufficient.

I think the analysis from the measure developers actually demonstrated that most of the impact analysis showed a 0.01 difference between the highest and the lowest performers. And one of the comments, the only comment that we got from the preliminary analysis which was very insightful said although the variation is hard to interpret with this 0.01 difference how significant is the performance.

And I think it really, again, it's the context question is it's really which lens you're looking at this and whether it's the hospital provider or whether it's the consumer looking at the Hospital Compare website or whether it's CMS looking at it. People get different I guess conclusions from the statistics.

And so if you're the consumer

looking at the value-based -- or the Hospital Compare website and you actually look at this number of -- you know, if you're a medical center and you're performing at 1.08 as the measure. And the median is 0.99, I mean what does that tell the consumer, right? So it tells them nothing.

But then if you look at some of the analysis looking at hospitals in the 90th percentile with a score -- costing Medicare 25 percent more than per episode than hospitals in the 10th percentile. You know, that's meaningful data to CMS. So I guess the criteria here is from where do you look at this data in terms of the lens and where you sit on the fence.

And in terms of if you're asking me, the provider, the hospital whether this is clinically relevant I personally would say it's of low relevancy. I think for CMS and for Congress this would probably have some difference but it's not statistically or

clinically meaningful.

DR. PENSON: Thank you. I appreciate that. So the last piece in validity is something we've already talked about a good bit which is -- I'm sorry, Lina, I didn't see your -- sorry.

DR. WALKER: I just wanted to follow up on that comment because that was my comment. And I have a question for the developer.

The score that's available to the public is just that single score which to me was meaningless, especially to a consumer who's looking at that score. It was impossible to interpret. If I'm trying to assess the performance of a hospital I wouldn't -- other than the up/down, above 1/below 1 it provided very little information.

But what you have said and what was available in the documentation is that the hospitals will receive additional information. And that information would parse out the

measure by different subcategories. That additional information would be meaningful to consumers.

And I don't know, I mean I guess my -- getting to the point of my question, you talk about intent in the beginning and I feel that this particular measure doesn't get to intent because there's so little information in the measure that's publicly reported. And if I were to evaluate it from the point of view of what's publicly available I would say, you know, it doesn't meet that criteria.

MR. ZAIDI: So that's well taken. It is true that on sort of the front page of the Hospital Compare website there is just one number. But the consumer can download files which show for every hospital the proportion of their costs in each of the seven types of service that I mentioned. So it's not posted right on the front page right now but that data is available to the public I think on data.medicare.gov. As well as the state and

national averages for all of those measures.

DR. WALKER: But is there a way to translate that measure though? Because I mean I think -- what you're saying is that it requires multiple steps to understand the measure. Is there a way to translate the measure if all you see is that index?

MS. SPALDING BUSH: I guess we're not sure what we want to translate it to. I mean I think it's a resource use measure, Medicare payment measure. So what are we looking to translate it to given that data?

DR. WALKER: Well, is a 0.01 difference a significant difference in resource use?

MS. SPALDING BUSH: Okay. So that's something you think the consumer would benefit from us explaining?

DR. WALKER: Well, we're working towards -- this is, as I understand it, a building block towards trying to get to value and efficiency. And so if I have no way of

interpreting the data I wouldn't know how to use that particular building block in conjunction with other quality measures to assess value and efficiency.

MS. SPALDING BUSH: I think that's correct if you look at the measure as it stands alone. A few people have mentioned the Hospital Value-based Purchasing Program which has the measure in a domain that's weighted in with the clinical process of care, patient experience and outcome measures. So then you get a total performance picture that way.

But it was developed as a standalone cost measure so that you can use it in conjunction with those quality measures to recognize a hospital that's providing higher quality care at a lower cost to Medicare.

But you're correct, I think standing alone it doesn't tell you quality. But I don't think that that was necessarily its intent.

DR. PENSON: Yes, so I think --

well, I think what you have to decide is you have to again look at the measure as it stands and decide if the differences you see are meaningful, whether statistically meaningful or clinically meaningful or just common sense. I mean they're there and I don't think there's any judgment to it. Daniel?

MR. WOLFSON: I actually think that it's to the advantage of everybody in the room that there is not a lot of variation between the players. So ultimately we can refine over time and nobody gets hurt in the middle or in the beginning. I think it's actually a good thing.

Two, the biggest unintended consequence that just came up which -- Americans will buy the most expensive item on the shelf.

(Laughter)

MR. WOLFSON: That's the thing that really scares me. Now, the Choosing Wisely campaign is going to change all of that

thinking.

(Laughter)

MR. WOLFSON: But I need a few more years. And that's the biggest unintended consequence, that when we put this out although things are changing because I think as insurance is involving the patient more in the decisions I still don't think they'll think about cost. And they'll want to go for the most expensive hospital. They don't like shopping at Walgreens for healthcare.

DR. PENSON: As a dedicated Apple user I don't know how you can say that.

MR. WOLFSON: I didn't hear you?

DR. PENSON: As a dedicated Apple user I don't know how you can say that Americans go for the most expensive things on the shelf. God knows we look for value.

So I will try to keep moving this along for the sake of time. Jack, question, comment?

DR. NEEDLEMAN: Yes. I actually

have a question for the developers but Daniel's comment, it's -- I actually think the issue here is not so much that people are going for the most expensive thing on the shelf. But they're -- consumers are very much afraid of skimping on valuable care. So it's the low-cost folks that make them wonder what am I giving up here. And so it's a different mind-set than the most expensive. It's am I being skimped on in these other places. And that's where I think the quality issues come in.

But Tom talked about the difference between the bottom 10 percent and the top 10 percent being 25 percent. And we had that other measure that Cheryl referred to earlier that when we looked at quintiles, the bottom 20 and top 20 percent, we saw a 6 percent shift in the hospitals in each one of those, 70 percent, 70 percent.

So, but if we're going to look at 25 percent differences now I'm concerned about

whether those are a function of the inherent variability of care. So if we go to the more extremes where we get this 25 percent difference can you tell us what proportion of the hospitals stay in the 10th -- the deciles, top and lower deciles? Or alternatively can you tell us what the cost difference is between the bottom quintile and the top quintile so we can compare apples and apples?

MR. ZAIDI: So I think you're getting at putting a confidence interval around the measure that we report.

DR. NEEDLEMAN: No, I'm talking about not so much a confidence interval but the stability of the measure. So we did test, retest and you know, at the top and bottom quintile 70 percent stayed in and 30 percent stayed out. That's 6 percent of the hospitals at the top and the bottom shifting at that point.

We heard in terms of how big are these differences at the decile level 10

percent and 10 percent, not 20 and 20, a 25 percent difference. But I'm wondering when you did the test-retest what percentage of the hospitals are -- stay in the decile. Because that will tell us something about how reliable, how stable this measure is of high cost and low cost.

MR. ZAIDI: So we didn't do a decile stability analysis. But I would say that that 30 percent that drops out of the top quintile, they might be moving just a little bit below the cutoff. So it's a very discontinuous measure.

And I think as I mentioned almost all of them are in the top two. So that 30 percent that moved out of the top quintile, they're almost all in the second most expensive quintile.

So a way of less discontinuous measure is to do the rank correlation. And that was, I think that was like 0.85.

DR. NEEDLEMAN: But what we're

hearing is fairly tight expenditure compression within the whole sample. So moving from the first quintile to the second quintile, particularly if you're right at the margin there may not be a very large expenditure difference. Or Medicare expenditure difference. So when we hear numbers being -- justifying this measure as the bottom decile to the top decile is a 25 percent difference in Medicare expenditures standardized. That's a bigger number than 2 or 3 or 4 percent differences which may be what those shifts at the quintile level are showing. So that's why I'm asking.

If we're going to think about the materiality of this measure which is part of what's -- part of the criterion, you know, how stable are the projections of the difference in spending from the low to the high.

MR. ZAIDI: Well, we don't report the quintile or the decile. Right now we just report the number and tell them whether it's

above or below the median.

But the broader point I would make about the materiality and the difference between. So the interquartile range is about 10 percent which translates to about $1,900 because the median is about $19,000.

But I would say that most of the variation between episodes has been gotten rid of by the risk adjustment. So that measure is actually very conservative because the risk adjustment is controlling for so many things that most of the variation between hospitals has been washed out by that procedure. So what's left I think you can have more confidence in the difference between hospitals that's left after risk adjustment.

DR. PENSON: So I think, Bill, I see you have a question. So I'm just going to ask folks to try to -- I mean I'm starting to hear problems with the differences but I will ask folks to so try to cover very new ground when possible only for the sake of time. So

I don't mean to cut you off, Bill, or stop you but I just wanted to remind -- I think we've established that some folks in the room have concerns with differences in significance. Bill.

DR. WEINTRAUB: Just quickly to remind everybody. There's 25 percent difference after risk adjustment. How much difference was there before?

MR. ZAIDI: I think it was about five times more. I don't have the number on me exactly. It was a much larger difference.

DR. PENSON: Okay. So other comments here? We sort of covered disparities before but we'll just go onto -- is that the next slide that you have, Evan? Okay.

So we've kind of gone through disparities earlier in the day. And we don't have to worry about comparability of multiple data sources. Brent, you have a comment? Okay, well no, we're not going to vote quite yet. So I just wanted to sort of summarize a

little bit and I also want to let anyone who has comments in the room make them and then go onto public comments.

So basically I've heard a bunch of different things as we've gone through this. I've heard some concerns about exclusion with regard to thresholds and the deaths. I've heard some concerns about adequacy of risk adjustment. And I've heard some concerns about the differences measured.

I've also from Carlos with regard to validity testing while he feels better about what's been presented he still has some concerns there. I'd like to invite additional comments for the overall validity piece and then we'll let the public comment and vote. So David and Brent, if you have one. Daniel, I don't know if you have a comment as well. David.

DR. GIFFORD: Well, I'm not sure where in the validity. I wanted to make some comments about their content, construct and

criterion validity testing. Is that later or is that right now? Okay.

DR. PENSON: It's right now.

DR. GIFFORD: So I think as we talked about their testing of the content validity with the expert panel and some of the other stuff was good, was okay. It has some face validity. But we've also talked a lot about today some concerns with the exclusions of Part D and other issues out there on the content validity of the measure and the other issues.

But I think most concerning is the thing we talked about, is the construct validity. The correlations with rehospitalization, multiple episodes when they add stuff in is really small with very little change. And there wasn't just in the writeup a thought process for what the direction should be that we'd see. I think we all intuitively know it but it's really poorly done. I mean it's not poorly done, sorry.

It's well thought through but I think there's a lot of other construct validity tests that would be interesting to look at.

They have some data in there but they didn't talk about it. They just sort of throw the associations out there. Urban versus rural, other types of hospitals would be interesting.

And then I don't know whether -- I'm trying to think of a criterion validity but it's probably really construct validity. But there are a number of all-payer databases out there and it would be very interesting to look at how this measure tracks with other cost measures of hospitals from commercial insurance or elsewhere. And that shouldn't be that difficult to do. It's not national but it's enough to give a good sense.

I think the results that are shown there are really weak and they're not even clear what the direction they should be, either divergent or convergent validity

testing. So to me of all the things we talked about today, I mean, I think everything was pretty reasonable and good. This is the one area that I had a really hard time in concert with all this other risk adjustment and everything we talked about. I mean most of it I could probably could get by but this just tipped me over to say it's really hard for me to vote on it with that.

DR. WALKER: Could you speak in plain English about what the concern is?

DR. GIFFORD: Helen's muted me I think. I had to go back to my epi days with Hal Morgenstern at the School of Public Health at UCLA. And actually we're trying to submit four measures to the NQF right now so we're having to go through all this.

You know, in thinking about -- and it was nicely summarized I think by Ashlie at the beginning. Your three tests of validity, the content validity, the face validity just don't make sense. And the way it tested is

you put experts together and if we all agree then that meets that criteria. And I think you're seeing that that's sort of iffy with everything we're talking about. Probably passes enough. Given the measures that are out there I think I would say that's okay. I'd like to see it better but I'd vote okay on that.

Construct validity is usually -- and there's other people in this room. Cheryl or other people, go ahead and correct me if I'm wrong. I didn't get an A in this, so. It really is that it should track either with measures that you expect it to track or not track with measures you expect it not to track.

So the hypothesis that's out there is that facilities that have lots of episodes of care or lots of rehospitalization should have higher cost than those that don't. And when they do the test they see a correlation but it's really, really weak. And is that

correlation strong enough for us to say we think that that passes the construct validity?

The issue of -- they show a lot of data showing that ethnicity really is related with higher cost in a lot of other studies. And when they stratify by the ethnicity and show the data and they adjust for it we don't see any correlation. That fails the validity test in my book.

You might say that there's data to suggest urban hospitals are different than rural hospitals. Do we see that trend there? Is it strong enough to say we think it passes the validity test? It might be that there's types of hospitals that are out there that, you know, women's hospitals we'd expect to see a very different cost structure. Does this measure show very big differences or not? We don't see the results of that. So I would like to have seen more of those types of tests.

And then criterion validity is the

gold standard. And I'm not sure we really have a gold standard so the closest I could come with a gold standard is we do have other cost structures. So there are many states that have all-payer databases which are claims submitted for not just hospitals but all settings from insurers. You could probably take one or two of those states and look at some similar type of measure and see if this measure tracks with that. It wouldn't be exactly correlated but it's just consistency.

So if Dave Gifford's hospital is really, really expensive on Medicare I'd expect United and Blue Cross to be really, really expensive too. And if it's not I'm going to go wow, what's wrong, why is this measure different. There may be legitimate reasons for it but I would think an a priori hypothesis would be that. And so I just would like to see more validity testing, more construct validity type testing here. I just didn't see it and that's my concern.

And in defense of the measure developers you never can do enough reliability and validity testing. You're often never given enough money by CMS or anyone else to do the testing for it out there. And no matter what you do someone can sit and be an armchair quarterback and criticize it and say that it's not enough testing out there.

So in that defense I don't expect it to be perfect but I would like to have seen a little bit more. And the testing that we did see I thought was very weak correlation and wasn't strong enough. That's my comment.

DR. PENSON: So thank you. Any other comments in the room first? All right. Hearing none we're going to open the floor to public comment either from behind me or on the telephone. So we'll wait and see if anyone has any comments.

DR. BANKOWITZ: Thank you, Mr. Chairman, for allowing me the opportunity to comment. My name is Richard Bankowitz. I'm

the chief medical officer of the Premier Healthcare Alliance which is an alliance of 2,800 hospitals throughout the U.S.

Let me preface my comments by saying Premier is working with all of our hospitals to drive down variation. I think we're all after the same thing. We are also working with 80 healthcare systems to create accountable care organizations which we believe is fully the way we need to go.

Having said that I want to address concerns with the measure we have before us which I think has been rightly identified as a measure of Medicare expenditure for hospitalization or Medicare expenditure for extended hospitalization if you want to look at it that way.

The first concern I have is with the unit of analysis. And I'm going to talk about the scientific validity.

You've been using terms like "the hospital has variation of X" or "we see a

variation of Y in these hospitals." The fact is the variation in the hospitals is zero. If you look at the CMS data in our fee-for-service world where we've got Part A and Part B expenditures the Part A expenditure variation is zero because the DRGs have cancelled that all out.

What's left is the Part B variation which is the professional fees and the ambulatory care, et cetera. We would love to have an entity accountable for that. It could be an ACO, it could be maybe a provider if we could identify an attesting -- attending physician. But we don't have one. So we've gone to the most convenient, the most convenient unit of measure which is the hospital.

And I don't think we would accept this if we were putting on our scientific hats and we were reviewing a paper and the author said well, we've chosen the unit of analysis mindful of the fact that it has nothing to do

with the variability we observed but it was the most convenient one to choose. We would not accept that. So I'm not sure why we accept it here. That's point number one.

Point number two is the socioeconomic status. Looking at the dual eligibles I understand that NQF does not want to mask disparity in socioeconomic status. But dual eligibles share more characteristics beyond socioeconomic status. We know they have usually multiple chronic illnesses, we know they have complex societal issues, we know there are disparities in healthcare literacy. And I don't think NQF says do not look at those things. They should be looked at. It would be as if we would say we're not going to look at diabetes because diabetes -- we might mask the disparities in care among the diabetic patients. We wouldn't take that approach. So we've got a population with chronic, complex disease and I think we need to account for that.

I would also say I take issue with the fact that there was no impact noted because if I look at the data that's presented on Table 14 in the appendix if you -- it's true that about 90 percent of the hospitals did not show any impact when you added that variable. But in 10 or almost 11 percent of the cases the impact brought the difference measure down by 1 to 3 percent.

So I guess you would say well, it's only 10 percent of the hospitals but that's maybe not important unless you're in that 10 percent. And I'd like to know who are those 10 percent. Are they safety net hospitals? Are they academic medical centers?

And moreover if you look at the data the bias is only in one direction. It only lowers the difference. It almost never raises the difference. So this is exactly what you would expect if there were a real impact of dual eligible. So I read the table differently.

The third thing I want to mention is using the MS-DRG in the regression and the particularities of doing that. You have to remember that you can get into a higher MS-DRG through a comorbidity or through a complication. So if patients have complications and they're bumped into the major CC you have now masked away the fact that they had something go wrong.

Even in an extreme case if a patient were to come in for a minor surgery and ended up on a ventilator through a complication for 96 hours they'd be in an entirely different DRG. So all of that quality gets masked. So we may need to look at DRG families, we may need to look at present-on-admission flags, but the way this is done masks any variation attributable to quality. So thank you very much for letting me comment.

MR. SHAW: Hi, I'm John Shaw from Next Wave in Albany. And I wanted to address

just two of the areas that seem to be the most problematic for the group. And that is the 90-day period to look for for risk adjustment for the HCCs. And the unexpected results that looking at the dual eligibles didn't seem to make as much of an impact.

And part of what I perceive as a contributor to both of those concerns are things that are not in there. They're exclusions that aren't really explicit exclusions, they're implicit exclusions.

So for the dual eligible population we're looking at a measure where we're measuring Medicare covered costs. The dual eligibles have Medicaid covered costs that are not in there. And the total costs for caring for them which we're used to seeing in terms of overall expenditures are both.

Yes, there's patterns where the dual eligibles are high on the Medicaid side and high on the Medicare side but without looking at all of the costs it's really not

complete.

Secondly, in looking at the diagnoses that are used for risk adjustment I was concerned too not looking out a year ahead because I know the small numbers of diagnoses that get reported and particularly in a 90-day period, the argument that those are going to be more relevant and closer to describing the episode is somewhat compelling except for one thing and that is the coding guidelines. The coders will only report on the encounter those diagnoses that affect the current episode of treatment.

So that means that if you're coming in and starting this with a hospital admission and the 90 days prior is all doctor visits and outpatient something is different. So I think you almost have to look at the index stay for anything that's prior -- present on admission to include in explaining what the course is likely to be for them. Thank you.

DR. PENSON: Thank you. Any comments from folks on the phone? Operator?

OPERATOR: At this time to make a comment please press *1 on your telephone keypad. There are no comments or questions.

DR. PENSON: Great, thank you. So I think we're at a point where we need to vote for the validity of the measure. We've had a long discussion. I'm not going to bother recapitulating it because I think you've all heard it. So let's go ahead and vote.

MR. WILLIAMSON: We will now vote on the overall validity. You will have 60 seconds. Please begin now.

DR. PENSON: Everyone try again, someone's missing. Oh no, I've got them all now.

MR. WILLIAMSON: We have 13 moderate, 11 low and 1 insufficient.

DR. PENSON: Well, there you go. Okay. So on that note -- actually, what does that mean?

DR. GIFFORD: Is that a statistically significant difference?

MS. WILBON: I'm tempted to defer to Lindsey who is actually working on consensus stuff right now. But essentially for the process, and I'll let Helen and Lindsey add to this, is essentially we would move onto evaluate the rest of the components of the measure.

It does pass scientific acceptability although I think there's probably some additional comments or some input. We could decide on whether or not that's actually consensus or not. But in terms of votes we would move forward based on the majority votes. Helen or Lindsey?

DR. BURSTIN: Yes, I mean one of the issues we've had, we've been spending a lot of time with the board level consensus task forces really identifying what's consensus. And I think, you know, I don't think there's a whole lot of comfort when it's

split like this.

I think at this point in the middle of a review you just need to have -- these are sort of a series of tollgates. So I think at this point you have at least moved past this tollgate, recognizing there's a lot of issues you probably need to return to before you get to the final decision-making. And then we'll have to see what the ultimate vote looks like when people overall assess how all the difference criteria come together.

But I think based on this you should just continue to move forward to the other criteria. Any questions?

DR. PENSON: So, yes. I think it's interesting to look at that and see that there were no highs there at all. And I think there is something of a consensus towards the middle to the bottom. And I guess as Helen pointed out it's really going to come out in the wash at the very end.

So let's keep moving along and

talk about feasibility. And I'm hoping that we'll be able to move through these last two points relatively quickly. I'm going to urge everyone, we've covered a lot of ground and sometimes we're covering it again and again and again. And I don't want to cut anyone off so please, don't beat the dead horse, just for the sake of time. Nancy?

DR. GARRETT: I just had a quick process question. So if it had been reversed and the majority were in the lower insufficient then what would that mean?

DR. BURSTIN: This is -- given how close it is we would probably still ask the committee to just finish the review. But technically this is a must-pass criterion and we try to respect that. Given that it's so close I think we just need to let it ride and finish it up. But it is a real concern. Unfortunately for some of the tough measures we see a fair amount of this. So let's just keep going.

DR. PENSON: So next we'll go through feasibility. And this is a number of different points.

As you can see in the discussion points here, it's really about the data being available, retrievable without burden, inaccuracies and can it be implemented for performance measurement. And because it's CMS data looking at the responses it was relatively well received there.

So Carolyn, you were the commenter on this.

MS. PARE: Yes, and first I'd like to thank staff for giving me this particular criteria because in comparison to everything else it's pretty straightforward and non-controversial.

I also think David pretty much covered the endpoint here in that basically everybody agrees that this is feasible. The data is being collected, it's available. I think the only concerns that I read and to do

a quick synopsis to save us some time is really the usual data issues of how clean is the data and how closely is the administrative claims data connected to clinical processes. Beyond that there's not a whole lot of input relative to this not being feasible because it's data that's already electronically available and collected.

DR. PENSON: Other comments from the room? I figured this one would be relatively easy. So I think on that if there are no other comments this one I think we can go to a vote fairly quickly. So Evan, go ahead.

MR. WILLIAMSON: We will now vote on feasibility. You will have 60 seconds. Please begin now. We have 23 high, 1 moderate, zero low and zero insufficient.

DR. PENSON: So the last piece is overall usability and use. And so this has got four pieces to it, accountability and transparency improvement, unintended

consequences and measure deconstruction. And I want to just, again for the sake of time, not to be rude that we have covered a lot of this already. So -- well no, we don't have to vote. We shouldn't go that quick. I'm not that hardcore, Joe, okay? Because I do think it's important to just sort of cover the landscape here. But I do want to remind people that we have talked a lot about the unintended consequences. And this is the place to reflect it in your vote but I think we remember what was said earlier. So with that in mind, Joe and Dolores, you were the two reviewers for this.

MS. YANAGIHARA: I will start out very briefly because like you said I think most of this has been covered. And then Joe will give -- I'll give more of the conceptual and Joe will give more of the real life kind of picture maybe.

So the first one on accountability and transparency. As the pre-vote shows this

is being used already so basically it meets that criteria, it's already in use for public reporting and will be in use for payment purposes.

Improvement. I think that most of the comments were it's a little too early to know whether it can really drive improvement and is there sufficient detail given to the hospitals to allow them to use it for improvement. I think that we've heard more today about the types of reports that are available and so it seems like there is a lot of detail that's given to the hospitals to be used.

That vote was a little bit more toward the medium/low than the high side, that is the pre-vote, but that may change based on the additional information today.

We've talked a lot about unintended consequences. I think it bothered people that it really wasn't addressed in the submission itself. Because people as we've

talked today have identified several unintended consequences that are possible. And the pre-vote reflects that.

In terms of measure deconstruction it seems like the comments were mostly around, more around consumers and could consumers really deconstruct the measure and understand it. And I think at this point probably not because the amount of information -- well, I shouldn't say that. The amount of information that's available on the website is clearly at a high level.

But if you are able to get the data to the extent a consumer would be able to download and understand and use the data apparently more data are available. But I think that's still a question in people's minds. So overall it's more toward the high and medium level in the pre-vote although there were definitely some concerns as we've talked about today.

DR. PENSON: Joe?

DR. STEPHANSKY: As you can imagine I have a few issues with usability. But first I might add based on a paper that I think Helen sent some of us by the American Association of Presidents of Statistical Societies, a very riveting paper which essentially -- the Hospital Compare is probably not the right name to use, maybe Hospital Obfuscation and so on. Because consumers aren't going to get anything out of a lot of these measures, not just this one. That particular paper had to do with readmissions more than with costs.

But the real question is are hospitals going to have the data they need to know who to talk to in the provider community and be able to assess the performance of some of these other providers. They may be able to reconstruct the number that is on Hospital Compare based on the data that you give them but in general they're going to need even more detailed data yet. We really need right down

to the patient level claim data, not just any aggregation of that.

But there's an unintended consequence there as well. Even if we had that data how many hospitals really have the analytic capacity to do anything with it? The pool of analytic talent, the big data guys, it's a very thin pool. So if we're assuming that hospitals are supposed to be able to use patient-level data to say regroup the patient-level data into episodes where they can look at care and particular clinical areas and determine where the high costs are and the high-cost post-acute care providers we're going to be making choices among hospitals who's going to survive and who's not. And I do think you need to think about that.

I think CMS and whoever they contract with needs to be ready to work with hospitals to actually interpret the data, not just set it out there and let them reconstruct the 1.08 or whatever score they earned. I

hope that makes some sense.

Our experience with the Blue Cross Blue Shield data is that just looking at the total cost by those cost buckets, SNFs or home healthcare and so on doesn't tell you very much because many of those post-acute care costs in that 30 days were actually related to other conditions that the patient had, not to that actual hospitalization.

We've been recombining the data into defined episodes to focus in on where the actual variation is happening within a clinical service line. And we're having much more success in identifying where we can cut costs and maintain quality.

DR. PENSON: Great. So I'm seeing names up. So Jennifer, you have a comment. Oh, I'm sorry. Yes, go ahead.

MR. ZAIDI: Yes, I just wanted to say that we do have an episode-level file. So it's not at the claim level. So every line is an episode.

And the cost buckets right now are the seven buckets that I mentioned earlier. So there are -- it's not going to tell you whether it's an X-ray or a CT scan. It'll say physician professional fees, or SNFs, or home health. And it will tell you the top five providers for each of those categories for each episode. So generally you'll know every SNF that treated that patient in that episode, or every home health agency that treated that patient.

So that information is available and we would be happy to -- it's an iterative process. So as we receive feedback from hospitals we have the ability to change those hospital-specific reports and the structure of that data. But that data is being made available to hospitals.

DR. STEPHANSKY: That makes a lot of difference because that was not described in the initial documentation as to what detail level the hospitals would have.

DR. PENSON: Great, thank you. That's helpful. So Jennifer?

MS. EAMES-HUFF: I just want to speak to the usability, to the consumer perspective in using this. And I would agree with what people say in terms of what gets posted on Hospital Compare has a relative limited usability to consumers without providing a lot more context.

I think the value of what's going out there is the downloadable files that you put on your website where many regional collaboratives and other organizations take that data and then report it in a way that is more meaningful to consumers. So I think there is ways that we can report this better.

And I don't have the expectation for CMS to do that. I've given up hope on that. That's my bias. But I do think there are others out there that can do it.

DR. PENSON: Brent?

DR. ASPLIN: Just a quick

question. And this is really in 4d, delta. You go from a ratio to an amount and then back to a ratio. And I just want to make sure that I understand the reason we're jumping back and forth there. Is that because the observed expected is based on the 25 different regressions you're doing and the various MDCs. And the amount allows you to adjust for the case mix of occurrences in the various different major diagnostic categories. Or am I off on that? I mean I'm just trying to understand why we've got to go from a ratio to an amount and back to a ratio.

MR. ZAIDI: So the amount is just the ratio, the observed to expected, just expressed as a dollar amount. So it's observed to expected multiplied by a national average. So you have some sense of what range of cost we're talking about.

So if your observed to expected ratio is 1.05 you just take that and multiply it by the national average. So that's the

amount.

And then the final measure is just dividing that amount by the median so that people have a sense of where the hospital ranks relative to the median. Does that?

DR. ASPLIN: Where do you go -- where do you -- because you're doing all the different, 25 different regressions, correct? So where do you adjust for the different case mix of cases in that to get to the single? Is that right as you go to the amount or where?

MR. ZAIDI: So, after the regressions are run you have all the episodes for the hospital. And you'll have a predicted amount and an actual amount.

DR. ASPLIN: So it's in that observed to expected.

MR. ZAIDI: Yes.

DR. PENSON: Dolores?

MS. YANAGIHARA: I just wanted to offer kind of a contrary view and maybe a little bit controversial view. But about how

much data the measurer should be expected to provide. It doesn't need to be all of the data that the people being provided need to be able to improve.

Because leading a statewide coalition in California that's voluntary and collaborative there's varying views from different stakeholders about how much data really need to be provided by the program versus that the groups or the entities being measured need to just be able to drill in and figure out on their own.

And so I think -- I don't want us to get down path of if not every detail needed by every hospital to be able to figure out exactly what to do to improve this is provided then it's not still a valuable measure. So I think that that's just something to keep in mind that where the responsibility lies to really understand the actual improvement.

Yes, it needs to point you toward where you may be an outlier and where you may

need to focus and where you may need to drill, but in terms of actually getting the member-level every detail that may or may not be needed. So that's just a --

DR. PENSON: Thanks. David, do you have a comment? No. Oh yes. No. Bill.

DR. WEINTRAUB: I want to speak a little bit about what we just heard on the last comment and also from Joe about how hospitals are going to respond, what are the unintended consequences, what's likely to happen.

So this is a classic big data problem. If you want to integrate everything that involves care in your hospital and outside your hospital and you want -- from a quality point of view and from a cost point of view to integrate those and understand them, understand how you're going to improve you need to be able to bring in all these data streams and analyze them. And Joe is right that in general hospitals do not have the

analytic capacity to handle this.

So what's going to happen? Either I think we're going to see that with things like this and many other forces in our society are driving us towards consolidation so we're going to have very large systems, the Kaisers of the world which have tremendous analytic capability and will be able to deal with things like this.

Kaiser's interesting because that's completely different in their payment model. But you see what I mean, a very large system with tremendous analytic capability. And I think we're seeing those just beginning to emerge.

How can hospitals respond? And it's true you don't have to have all of this, all of the data capabilities to respond a little bit. But to fully respond to the forces in our society will ultimately require those kind of data capabilities. As I often say the healthcare systems that are going to

thrive are the ones that can handle their information. I've never heard anybody disagree with that.

DR. PENSON: Thank you. Joe?

DR. STEPHANSKY: Just one more piece related to that in terms of what information I need. Because I need to know who to talk to in my community if I don't own them or otherwise control them in a way to affect their behavior. Let's say the SNFs. You may tell me who the top five SNFs are that I'm discharging to but what I need to know is which of those SNFs on the Medicare side always use up the full Medicare days and which ones are only holding those patients as long as they need to be there and discharge them from a SNF.

Unless I know that I don't know which ones to talk to and which ones to try and partner informally with. It's key that I know the actual provider.

I'm going to go back to the Blue

Cross Blue Shield piece again where the information being shared between the physician groups and the hospitals, I know as a hospital which physician groups at Blue Cross Blue Shield are sending their patients to me. And I know what kind of costs are happening afterwards. I know which providers in the community to try and partner with. And that's going to be a key here, not just a list. I need to know the costs too.

MS. SPALDING BUSH: I think that's a good point and it's well taken. I think what you can see at the episode level is the SNF cost for each episode. So every discharge that creates an episode you can see the SNF cost and you can see the top five SNF providers that treated during that episode. So if your beneficiary went to more than one SNF you may have two in there. You would have a pretty good sense I think if you looked at what were your really high-cost episodes and who were the SNF providers that treated them

during that episode. Is that getting at?

DR. STEPHANSKY: The actual providers, not just a dollar amount.

MS. SPALDING BUSH: Right. Correct.

DR. STEPHANSKY: Okay, well that's more detail than I thought was in there. We shall see.

DR. PENSON: Great. David Redfearn, then David Gifford.

DR. REDFEARN: Just a question maybe for NQF. We're talking about usability. Usability to who? And there's two we've talked about. One is members and the other is hospitals.

My conclusion is it's not very usable for members. It's potentially usable for hospitals. So are we evaluating both types of usability? And if we are, how are they weighted in our minds in terms of making the decision about overall usability?

MS. WILBON: So actually probably

both. The way that we phrase it here is that the measure can be deconstructed for those being measured and potentially also for those using the measure. So you can drop in those buckets whoever that may be. So it may be those using the measures, it may be purchasers and consumers.

Those being measured obviously would be, for this particular measure would be the providers, or there may be other -- the providers as in hospitals or physicians. You could maybe put them both in that bucket. But again I think this is another one where each of you are going to have to weigh, you know, how important those things are depending on the perspective and how usable you think it is or transparent you think it is for those particular groups. So it's a value choice I think for each of you again.

DR. BURSTIN: Although just one comment on that is that usually the measure developer has to indicate the intended

audiences. So the way we have our usability criterion clearly says what's the extent to which potential audiences are using or could use performance results for both accountability and performance improvement. So some of that really comes back to who is the intended audience.

And it's both -- especially for new measures. It's not are they using it because we can't actually say that, but it's also could use that information for both -- yes, for accountability and improvement.

DR. PENSON: David?

DR. GIFFORD: I'd just say that I mean every measure I've seen in every setting, whether it's a quality measure or this has unintended consequences with it. And each one of them is possibly gamed. And I think it has to balance it in the end.

But I think what we're hearing and Dave sort of said it, this would be useful -- this will force hospitals to look at what's

underneath the covers and what goes into this and begin to look at it. It may not be -- all be robust as Kaiser or someone else but they will begin to look at this.

And actually transferring back and forth while it could happen it's kind of hard. If you're balancing with other measures which we've all said has to be there I think we get there. So I think -- I find this measure in the sort of cornucopia of measures that are out there incredibly useful and usable and has no worse unintended consequences than any other measure that's out there.

DR. PENSON: Martin?

DR. MARCINIAK: So in just listening to some of the conversation, you know, we've touched on a number of different topics. The thing that just comes out to me and I'd be curious to hear from the measure developers is along the lines of the question of interpretation one of the things that struck me when we were starting to click

people off in terms of that high cost/low cost is really trying to understand what's driving those costs and actually driving the cost throughout the system of patients who are going through this in the hospitals.

One point we had made was the fact when these are put on web tools people see it, may see it out of context. Is there the opportunity to help people understand where the drivers of the cost are within the measure as we see it today and can that actually be reported out?

MS. SPALDING BUSH: So I think if I'm understanding the question it's about that -- so on Hospital Compare we only display the one number, that's the ratio. And that website is designed to sort of be a simple, understandable website for consumers to access. So we were kind of limited in what we could actually display there.

That's where you would find the link though to the -- you'd have to be

computer savvy and click on a link and go there. But you can find a link to the spending breakdown by claim type that has each provider's spending amount broken down and compared to that provider's state and to the nation. So that information is there and accessible within two clicks maybe.

And I think if there were other useful files we could do something similar, link from there to a data.gov site in the future. But that's all that's linked there now.

DR. PENSON: Terrific. So I think at this point I don't see any new flags up. David, I think -- you don't have a comment, do you? That was from before. Okay, just want to make sure everyone's heard.

So I think then the next thing to do is to vote on the usability and use. And then go onto a final discussion. So Evan?

MR. WILLIAMSON: At this time we will vote on usability and use. You will have

60 seconds. Begin now. And we have 6 high, 15 moderate, 3 low and 1 insufficient.

DR. PENSON: Okay. So we're getting towards the end with this one. We're getting to the point where we're going to sort of take an overall vote on NQF criteria for endorsement and whether or not it -- does the measure meet those criteria.

I think we'll start with any remaining discussion in the room and then I think one last shot at public comment. So I'll open the floor to parting blows as it were to this one. Dr. Gifford.

DR. GIFFORD: Having just reviewed two manuscripts in the last week I feel like the same recommendation to the editor. I really like the topic of the article. They're really addressing an important issue. I think it needs to be out there. I really encourage them to make some revisions and resubmit. But it needs some submissions. But they definitely should resubmit and it needs an

editorial. That's kind of where I'm leaning.

But I'm really worried that, you know, sometimes when you do that waiting for perfect can be the enemy of the good. So I'm actually kind of looking forward to a broader discussion of the group. Because I'm sitting on the fence as to which way I should vote on this. Because I think there's a lot of really good things about this measure but I think we also identified a lot of opportunities for improvement.

And if they took them back and came back with some of those I think this would be a very powerful measure and go through very quickly. But the question is do we let it go as is or not. And I will be swayed by the discussion.

DR. PENSON: So I want to influence the discussion a little bit not necessarily for the sake of time but for the actual process. I think depending on what the final vote is here I think the measure

developers have heard many of the concerns. So I think we have to vote on what we have in front of us. We know how it can be better but the question is is it good enough.

And I think even if it's good enough I'm sure that the measure developers will consider if it were approved as it were used, you know, tweaking it even further. But the bottom line is as it's in front of us we're going to have to make a decision. So I just want to stress that to folks. I think David's point is very important, very well taken. The key is as it is now is it good for go.

So we'll just go around the table. Bill?

DR. WEINTRAUB: So is it good enough. Probably yes, but just barely. And where does the concern lie?

So it's important, it did well on reliability, it did pretty well on usability. But really the concern seems to be in

validity. That is are we measuring what we think we're measuring. And we just barely had more people in moderate than in low.

I'll say publicly I voted for moderate but I'm concerned about so many people voting low and passing the measure. Where this is really critical is it measuring what we think it's measuring?

DR. WALKER: I echo what David and Bill just said. But I'm actually -- so I had -- my concerns were around the validity. And initially I had in the preliminary assessment ranked it low. But I was persuaded by some of the discussion at this table and had voted moderate after hearing the discussion.

And I'm on the fence of how to vote on this for the overall measure. And I'm looking to the technical experts at this table to help me connect the dots.

And so my question is one of the issues around the validity seemed to be that there was very little correlation between some

of the indicators of disparities and that seemed to raise some concerns. And there was very little variation in sort of the performance measure. And so there was a question of how -- are we over-controlling.

But our developers had I thought provided a very compelling, persuasive explanation of why you wouldn't expect to see some of that correlation, or a low level of correlation with these indicators of disparities.

And in fact when you gave us the data point on what the variation was between the top and the bottom quintiles without the risk adjustment it was huge. So it seems like the risk adjustment is absorbing a lot of the variation in the inpatient setting.

And then as Jack was saying most of the variation seems to be picking up all the post-acute care cost. So of course we're not going to see those correlations given what's happening.

So my question then is is it still a concern from the validity perspective that we're not seeing the types of correlation in health disparities as we would have expected. I mean, perhaps they're over-controlling for differences in -- perhaps their control is picking up some of the differences in socioeconomic status and maybe the dual status. But some of us had recommended or would prefer that we see controls for SES.

So I don't know how to evaluate it and I'm looking to our more technically savvy colleagues to maybe help me connect the dots a little bit. I mean is it still a concern?

DR. PENSON: Maybe some of the other comments will answer that for you I hope. Andrew?

DR. RYAN: The comment or question I have is what exactly are we voting for here. Because does CMS have to implement this measure kind of as specified now, or could they make some -- we approve this measure but

then there's some tweaks along the edges. Because I've seen, for instance, for the mortality and readmission measures that were once approved by NQF the risk adjustment has kind of been tweaked over time. And so these measures are in some way alive. And it seems to me that there is an ability for CMS even, you know, we approve it but they could still have a continuous improvement process. So I just wanted to get a sense of what happens and what it means to actually approve.

DR. BURSTIN: That's a great question. So in general there are sort of minor tweaks done along the edges. AHRQ and CMS know this, coding and things like that that don't rise to the level of being material or significant changes that they can just do. We see those on annual updates.

When there's a material change that actually affects the measure or the measure results we actually will do what's called an ad hoc review which we just did to

update the planned readmission model for the readmission measure. So it's true, the measures are alive, they can be modified, they can be changed, improved over time. And if there's significant enough changes they can come rapidly back through NQF for re-approval.

DR. RYAN: For instance, changing how transfers are treated, would that rise to the level of a subsequent ad hoc review?

DR. BURSTIN: Yes. And the only question is, and I don't know what the degrees of freedom are for CMS to actually potentially even, you know, depending on how this all goes, I don't know if they have an appetite to potentially even make some of those minor modifications during this process. But that's a bigger issue.

DR. PENSON: David?

DR. GIFFORD: I've chaired a couple of these and we voted a few down to CMS. And it's ranged anywhere from we never see it again to it comes back really fast from

CMS. I'm actually on two CMS TEPs that are reviewing measures that were voted down by NQF to try to get them back through. So you can see it back.

And then also they are tweaking it so they'll put TEPs together and it just depends on what it is. Sometimes they'll -- if they're under pressure from Congress they'll put this out and then bring it back because they still want to get NQF endorsement but they have to meet a congressional mandate so it doesn't really matter what NQF says. So it could be all over the map.

But again we're also not approving this just for CMS. We're approving it for anyone else out there. I mean, there's no reason that United couldn't go and calculate this on all the hospitals out there. This is all claims data, it's out there. It's no reason that a state couldn't figure out how to do this with claims data and measure it. So we're measuring a measure that's going to be

able to be used out there for all sorts of reporting processes.

DR. PENSON: You know, I think you know again the concept is as the measure stands now. As you see it in front of you can you live with it. If the answer is it's not perfect but I can live with it and I hope CMS makes improvements I'm going to have some faith in CMS and say you all will do that.

If you look at it and you say this is just not ready for prime time right now then I think you have to vote it as it is now with the understanding that if it's imperfect I think they've heard what people are saying and they'll continue to tweak it. Jennifer?

MS. EAMES-HUFF: I would say I don't think it's perfect but I think it's worthy of enough to pass a vote.

I think one of the things that I'd consider even though it does need some tweaking in looking at it is how they're planning on using it and what the impact will

that be.

And so they've indicated it's going to go in the IQR program which is not a performance program. That's a reporting program. So hospitals will not be getting their payment adjustment based on the actual performance scores. It's just are they doing -- is it being measured. And of course it's CMS using the claims data so it will be measured.

And then when it moves into the Hospital Value-based Purchasing Program they have three composites and the third composite is the cost composite which makes up 20 percent of the total score. So again I think -- when I look at this I think there's a concern around the financial impact and the potential unintended consequences that could come from it. It seems like that that is less than the impact of the attention that this will draw and the focus this will put on improving an area where we don't have a lot of

measures.

DR. PENSON: Great, thank you. Larry?

MR. BECKER: Thank you. So I'm going to state what I think I've heard and what conclusions I draw from that.

So I'm sorry, I don't remember your name but at the beginning I think you stated that the law calls for the development of this measure, if that's accurate. We are at an inflection point because it seems to me based on what I've heard is that we could turn this down but CMS is going to have this measure.

And so I think we're incredibly privileged to be able to have this amazing group of people together to comment, to give constructive feedback around this measure to make it better. But I think if we turn our back on this measure it could be, and somebody correct me if I'm wrong, but it could be that CMS is just going to go ahead with what it

wants to do without us. And what does that portend for the future?

DR. PENSON: Well, no, I mean it's out there but the question is, and I think the comment that David made is important. With NQF endorsement you'll see potentially uptake by private payers. And certainly there's a sort of imprimatur that comes with that endorsement. So even if by law it's going to be used by CMS I think there's something to getting the NQF endorsement that will make it more widely accepted.

So I don't think that people should vote yes just because the law says it's going to happen. I think you vote yes if you believe it's adequate. If that's fair.

Lisa? Okay, any public comment from the back of the room or on the phone? There's more in the room, I'm sorry. Who else? Hi Nancy, sorry.

DR. GARRETT: So I think for me what I'm really thinking about right now is

that validity part. And to see such a split vote on validity which is a pass or fail criterion for the whole deal, that's a real concern. And so I feel like I have to separate the uses of this and the fact that it's written into statute and all of that. And it's really about the measure before us. And those validity concerns I think are really big. And so that's a real issue I think we all have to think about as we make our decision.

DR. PENSON: Other comments in the room? Bill?

DR. WEINTRAUB: There may also be variation in how CMS uses this with and without endorsement. They're going to be far more inclined to use it for Value-based Purchasing, for instance, with endorsement than without one would think. No? Important point.

DR. PENSON: Tom?

DR. TSANG: So I have a question

to CMS and to the measure developer actually about the lag time of the data. Historically the claims data has always been at least, I don't know, having worked with CMS in my past life historically it's always been a question of a year to a 2-year delay in terms of getting some of the claims data back to some of the provider organizations and some of the quality organizations. So I'm wondering is this also the case for this measure in that if it's going to be a year to a year and a half to a 2-year delay or a lag, let's not call it a delay, then number one, how useful is this information giving feedback back to the hospitals if it's not within a relatively timely fashion?

MR. ZAIDI: So I think things have improved recently in terms of the claims delay. So for the period of performance for 2012 we plan to post those numbers in October of this year on the Hospital Compare website. So that's about a 10-month delay.

As far as the payment adjustment for Value-based Purchasing that score is based on two periods of performance, a baseline and then a performance period in order to measure improvement and achievement. So there is a delay in the translation of your score into your payment, but the information itself is visible within a year.

DR. PENSON: Great. Dolores?

MS. YANAGIHARA: Sorry, not to drag this on but a question about endorsement. So if this is endorsed would it be endorsed for all populations or only senior populations? Because I think the risk adjustment methodology has been calibrated to the senior population. So I'm just a little concerned.

I've been ruminating on what David said about it's out there and commercial payers will start using it. I'm like, hmm. So I just want to make sure what we would be endorsing, for senior population or any

population?

DR. BURSTIN: Currently it would be endorsed for whatever the population that it's been tested and assessed on.

As an example though in the past CMS has previously brought us some of the mortality measures just on 65 up. We had asked them to try to make that a broader population. They had to go back and reassess the risk adjustment methodology and see if in fact it works for those under 65. So this is actually just 65 and over but it does potentially give an opening to think about how those data could be adjusted to make it work for the broader population as well.

DR. PENSON: Great. Other comments in the room? Okay. Public comment either behind me or on the phone? We have a comment.

MS. WHEATLEY: Hi, Mary Wheatley from the Association of American Medical Colleges.

And I just wanted to thank everyone for the discussion today. It was a really thoughtful discussion about all the issues related to this measure. And as it comes out this is a really complex, all this resource measurement is very complex and how it gets implemented is very complex.

I would just like to say something about how the data can be used. And I think a general question for the group here is what is improvement in this measure. And maybe this goes into a MAP discussion down the road is how do you monitor improvement.

Because although we got new data I think a lot of hospitals are trying to figure out how to use that data. I think it might help in a few cases identify extreme outliers and ways you might partner in a few cases to reduce some of the costs.

But is it enough information for someone to easily go from a 1 to a 0.99. And what does that mean? I think people are

really struggling with how do you implement this, how do you take this information you have into something I can implement into an operational piece. So I just wanted to make sure that piece was out there.

And again, the whole question of the scoring. I think that is also something that would be a good thing to bring up in a MAP conversation outside of the scope of this measure. But there are many ways that this measure gets pulled throughout the whole process. Thank you.

MS. CHAMBERS: Hi, I'm Jayne Hart Chambers with the Federation of American Hospitals. And I could just say ditto.

But I thank you for your robust discussion this morning. It was very interesting. I think our members are a little troubled by some of the discussion around reliability and validity of the test data. We've had some of those discussions internally.

I will say in terms of usability it's difficult to take the information and be able to parse it in a way -- even if it's only year-old data to be able to parse it in a way that can help us figure out what our relationships are with a number of the post-acute providers that our patients would go to.

And there is that element of patient choice where they choose where they want to go and how they use that. So we are concerned about the usability of the measure.

And I very much echo what Mary said in terms of how it's being used in Value-based Purchasing and how you can drive improvement with it. Where you see a lot of compression with the scores and how do you get from 0.99 to 0.98 or whatever it is that you need to be driving down to when you get to that level where you have an achievement score and an improvement score and you want to be able to earn points on them. So thank you.

DR. PENSON: Thank you very much

for the comments in the room. Any comments on the phone?

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

DR. PENSON: Okay, so I think we're at the point now where we make a final recommendation. This is a simple yes/no vote. So Evan?

MR. WILLIAMSON: At this time we will vote on the overall suitability for endorsement. You will have 60 seconds. Please begin now. We have 15 yes and 10 no.

DR. PENSON: Okay. So Gene, I think it's your turn now but probably people need a break. Is that a fair statement? I kind of figured that. Me too. How long do you want to take? Dr. Nelson, when do you want to start?

(Whereupon, the foregoing matter went off the record at 2:30 p.m. and went back

on the record at 2:48 p.m.)

DR. NELSON: Okay, well I'd like to congratulate you for getting through the first measure. And it was a really good discussion. And I think you all did a great job. And I don't know if you knew this but David was on a redeye flight and so he's been up pretty much all night and did just a great job of facilitating this. And stayed awake the whole time. So thank you, David.

DR. PENSON: So now I'll be snoring for the second measure.

(Laughter)

DR. NELSON: You did well. Well, our second measure then is number 2165, Payment Standardized Total Per Capita Cost Measure for Medicare Fee-for-service. And the developers are CMS and Mathematica.

The measure that we just looked at was the hospital episode for Medicare beneficiaries. So that was being stretched out. And this measure of course is stretched

out even more. So it covers a year of time and is meant to focus on providers, on group practices specifically of a substantial size. And we'll discuss that as we proceed.

But similar to the prior measure it includes Medicare Part A and Part B expenditures plus some other expenditures that we'll hear about. So it's a little bit expanded in terms of the services that are covered.

It is price-standardized. It is risk-adjusted. And the way that the patients become associated with a group practice is through attribution. And we'll hear more about that. But essentially it's a plurality of primary care services provided by a particular group relative to other groups elsewhere.

And so having said a few words about the measure we should ask the measure developers to give us a brief presentation. You're over there and not here. Okay, so you

can introduce yourselves and get started. Thank you.

MR. BALLOU: Good afternoon. I'm Jeff Ballou with Mathematica policy research.

DR. RICH: And I'm Gene Rich with Mathematica.

DR. ROMAN: Sheila Roman with CMS.

DR. KAPLAN: Frank Kaplan, CMS.

MR. BALLOU: So I'm pleased to introduce the Payment Standardized Total Per Capita Cost Measure for Medicare Fee-for-service Beneficiaries. This is on the next slide a per capita resource use measure and mindful of the distinction that folks make between costs, expenditures and resource use. What we're really talking about here is service utilization on a standardized basis.

It's a measure that is applied to Medicare fee-for-service beneficiaries attributed to medical group practices. It includes all Medicare Part A and Part B covered services during a calendar year.

It does not include Part D pharmacy data. As was indicated this morning there's a significant proportion of Medicare fee-for-service beneficiaries who are not enrolled in Part D. And so including those data in this measure would unfairly disadvantage those groups who would happen to be attributed a disproportionate share of Part D enrollees.

The measure was developed specifically for Medicare fee-for-service beneficiaries and the methodology has been tailored accordingly. So, for example, the payment standardization algorithm which is the same algorithm that was used in the measure discussed just previously is specific to the nuances of Medicare fee-for-service claims and payment systems. And likewise, the CMS HCC risk adjustment model which is foundational to this measure was developed, tested and calibrated on Medicare fee-for-service claims.

And then finally by way of

overview this is the only per capita resource use measure among those either endorsed or under consideration for endorsement that is developed for and tested on this older Medicare fee-for-service population.

On the next slide in terms of importance and use we see the measure as a comprehensive tool that can be used to document and address variation in resource use among Medicare fee-for-service beneficiaries. It is a broadly inclusive whole patient measure that will complement additional, more targeted resource use measures.

It is included in confidential physician feedback reports that we provide medical group practices. And to identify drivers in resource use and variation from the national mean we report the measure not standalone but along with service-level detail, breakdowns of cost, for example, hospitalizations and the standardized risk-adjusted cost associated with those even

including detail on which hospitals are most frequently used by the beneficiaries attributed to the group.

We will soon begin -- meaning with this upcoming cycle of feedback reports that will go out later this summer -- begin also including beneficiary-level detail in the reporting tool so that providers will be able to see which beneficiaries were attributed to them and the characteristics and cost-related data of those beneficiaries.

In the future this will be included in the value-based modifier that will adjust physician fee schedule payments beginning in 2015 for selected medical group practices. Again here it will not be used as a standalone measure but will be combined with not only quality measures including outcome measures but other resource use measures as well to arrive at an assessment of value.

On the final slide I'd just like to give a high-level refresher of the

construction of the measure. From our target population we exclude beneficiaries for whom costs are unlikely to be adequately captured by Medicare fee-for-service claims.

We then attribute beneficiaries to group practices based on the plurality of evaluation and management services using codes that are commonly billed by primary care providers.

We then standardize attributed beneficiaries' Part A and Part B, just to be explicit here, Medicare-allowed charges. So we're standardizing Medicare payments as well as coinsurance deductibles and so on.

And then after handling outliers we use those standardized numbers to compute per capita observed resource use for each group. We then likewise compute or estimate per capita expected resource use as a function of beneficiary risk for each group.

And to clarify what might have caused some confusion in our submission we

used risk scores that are produced by the CMS HCC beneficiary model. And again that's foundational to what we do.

But then we take those risk scores and we regress our payment-standardized costs on essentially those scores to better fit the risk scores to our data. And then we use the coefficients from that model to estimate per capita expected resource use.

We then compute the ratio of observed to expected and express this in dollar terms much as the previous measure does by multiplying by the unadjusted national mean. And then comparing to that to the benchmark mean so that scores below the mean represent better performance.

Obviously this is a very quick and high-level overview. We are happy to entertain any questions of clarification as the steering committee proceeds. Thank you.

DR. NELSON: Thank you for that very instructive overview. Questions for the

measure developers? Daniel.

MR. WOLFSON: I would like a better explanation of attribution, whether that attribution technique has been used previously and tested. I've never seen that attribution. I've always been used to visits and not costs. So could you kind of give a history of the attribution technique and its testing and so on?

MR. BALLOU: Certainly. I can provide some information and my colleagues from CMS may also have information to add.

To the first point about counting up services, E&M visits, for example, office visits as opposed to the charges associated with those. We have tested both of those rules in a variety of settings. We find in general that attribution is very highly correlated using those two different approaches.

The approaches based on charges have been used in other CMS programs

previously including the PGP demonstration and more recently the group practice reporting option of the Physician Quality Reporting System. So attribution based on charges has been used and tested and accepted in those initiatives previously.

In terms of this two-step rule it is true that implementation in this measure of the two-step rule is new as of this cycle. What is meant by the two-step rule is it's meant to be a primary care-focused rule.

So first of all, in the first step you're looking for the plurality of primary care-related services that were billed by primary care providers. I'm sorry, primary care providers by which I mean physicians, namely those with the specialty of internal medicine, geriatric medicine, family practice or general practice. That's correct. So you're looking for -- and that's an important qualification, thank you.

We do, however, recognize that in

some cases specialists play the role of the primary care physician for some patients. And in those instances where there has been no primary care physician involvement at all we consider attribution to those specialists. And that's the second step.

This is a rule that has been implemented by the ACO initiative, Medicare Shared Savings Program. And the reports they currently receive which are updated on a quarterly basis use that rule.

DR. NELSON: And the description of services that are primary care services that get attributed would be?

MR. BALLOU: Actually, why don't I turn to my colleague Dr. Rich who may be in the best position to answer this.

DR. RICH: Sure. The specific history of the selection of those primary care services our colleagues from CMS may want to refer to. But they are largely ambulatory care or home visit or skilled nursing facility

visits which are -- many of which of course are done by primary care clinicians or primary care physicians.

So they exclude the very -- the services, the evaluation management services that might be provided in inpatient settings as well as excluding some other services, E&M services more typically provided by specialists or in some cases not covered under the Medicare fee-for-service program.

DR. NELSON: I'll start down the table. So, Brent.

DR. ASPLIN: Since we're on attribution what percentage of overall eligible beneficiaries actually end up being attributed by the method that you're proposing? What percent fall out? And of those that are attributed what percent actually are attributed based on the primary care step versus the specialty care step?

MR. BALLOU: Right. So I don't have the numbers exactly in front of me for

the first question. I recall that it's on the order of 90 percent attributed but I would have to check that.

Among those who are attributed, however, 85 percent are attributed in the first step and 15 percent are attributed in the second step.

DR. ASPLIN: That's a really high percentage of overall attributed potential beneficiaries. It's just really surprisingly high. I mean I know it's Medicare, not commercial, but I'm surprised that high of a percentage attribute which makes me have more questions about the attribution model.

DR. NELSON: As we go further into this discussion on all the criteria we'll get into many of these points. So why don't we take a couple of more general questions for the presenters and then we'll be able to get into the details as we move forward with the criteria. I think you were next.

DR. MCHUGH: So the first point in

the two-step process is a primary care physician claim trigger. Has there been consideration of non-physician? What happens to those beneficiaries who wouldn't be captured under that trigger but are still Medicare beneficiaries receiving primary care services by non-physicians?

MR. BALLOU: If they don't receive any primary care services from a physician they are not attributed under this rule.

DR. MCHUGH: But if they do receive care by an eligible provider who is not a physician they are excluded.

MR. BALLOU: Not necessarily. As long as they received some services from a physician in the group that is the candidate to which they could be attributed. It could be that most of that care is being provided by a nurse practitioner, for example.

DR. MCHUGH: But it has -- but the trigger is a physician.

MR. BALLOU: The trigger is a

physician in that respect, yes.

DR. MCHUGH: Is there a rationale for that?

MR. BALLOU: I think the best rationale that I could give is that this is a program that's meant to ultimately provide feedback to physicians and ultimately to grade physicians on their performance.

DR. MCHUGH: When you say feedback to the physicians are you talking about through the quality reporting mechanism?

MR. BALLOU: Feedback in terms of the reports that we give them and then ultimately they will be the ones who are judged in the value-based payment modifier.

DR. MCHUGH: But aren't more than physicians considered eligible professionals under that?

MR. BALLOU: Yes, that's correct. But they are ultimately not the ones being evaluated. The eligible professionals enter here because we're looking for groups of a

certain size which may include and often do include more than just physicians. But ultimately it's the physicians who are impacted in this particular programmatic context.

DR. MCHUGH: Would you stipulate that there's the possibility that you exclude some Medicare beneficiaries who are receiving primary care services under that attribution?

MR. BALLOU: Yes, I think that's true.

DR. MCHUGH: Okay.

DR. NELSON: Let's see, Cheryl? And then I think rather than keeping going we'll go into specific criteria and continue asking the specific questions.

DR. DAMBERG: Okay, great. Thanks. I was curious, the thinking behind using the prior year's HCC score versus something that's concurrent.

MR. BALLOU: There are a couple of issues here and our colleague Greg Pope from

RTI International I believe is also on the phone. And he is our expert having worked for many years with the CMS HCC model. So he may have a response after mine.

But I think the high-level thinking is that there was a concern that although concurrent models generally explain much better in large part because they're picking up a lot of the acute conditions that don't necessarily show up in the prior year.

There is a concern that it would be easier to confound treatment and diagnosis information in the concurrent model. There's also the problem of observational intensity bias that folks have been writing about and reading about lately that we believe is intensified by using the concurrent model.

I'm not sure if Greg would have anything to add to that though.

DR. NELSON: Okay, well thank you for those opening comments. You're going to get many questions.

And our first set of criteria have to do with importance. And we're looking at some discussion points so our large numbers of people, patients affected by the measure as well as providers in this case.

Does the measure demonstrate variation in resource use or overall poor performance? Are there patient or societal consequences of high or low resource use? And opportunity for improvement. Do the data demonstrate a distribution of performance scores?

Is the number and representativeness of the entities included in the measure performance data? Are there data showing disparities in the use of resources or cost of care for certain populations? What's the size of the population at risk and potential consequences of the cost/resource use problem?

And criteria 1c is measure intent. So is the intent of the resource use measure

clearly described. Is the construction of the resource use measure consistent with conceptual construct and the purpose of the measure? Do the resource use categories specified align with the intent of the measure? Are all of the categories captured in the measure that you would expect based on the measure intent?

And to get us going on this importance area we'd like to hear from Jennifer and then Carolyn.

MS. EAMES-HUFF: Okay. And in the interest of time Carolyn and I agreed that just one of us is going to present. So -- and also in the interest of time since the scores are pretty high on this I'm just going to keep this really quick.

So, overall in terms of the importance to measure and report everybody but one person gave this a high score for the high impact. In the opportunity for improvement still a significant amount scoring this high

but there's a little bit more variation with -- it's 16 to 4 giving it a medium and 1 giving it a low.

On the measure intent this is where it had the most variation in scores. And the high was 14, the medium was 5 and the low was 2.

In looking over the comments, at least the comments on the high impact a lot of them just said it's an important area to measure, it's a part of affordability and the Triple Aim. There's variation. So there seemed to be an overwhelming consensus on the importance of it.

For the other two areas there are a couple of comments that I'll just highlight. I think a lot of the comments that came in this particular area actually apply to other areas so I may just list them. But I'm not sure they're necessarily open for discussion at this point.

So the issue raised in opportunity

for improvement was the issue that this is an aggregate measure, that this is looking at total cost of care, that it's not broken down more discretely. So the question around the ability to do improvements and drive improvements was the main issue.

And on the measure intent the issue that was under there was around the reporting. And this is similar to what we saw in the earlier discussion of is this being reported on its own or is it going to be tied to quality measures.

There are a few other issues that were raised that I'm sure we'll talk about later. The issue of attribution and also again the issue of definition, of is this efficiency, is this resource use, is this expenditures. That was also raised in this as well.

DR. NELSON: Thanks, Jennifer. Let's have further questions and comments. Jack?

DR. NEEDLEMAN: Just a real quick question which was raised by the comment from the panel. I think it's important. I don't think that's going to be an issue.

But I just want to get clarification on the unit of analysis here. Because as I read it it looked like it was the group. And what I heard a few minutes ago said we expect this to be allocated down to the individual clinician. Those are two different levels.

And I just want clarification whether we're looking at a group-level measure or an individual clinician-level measure.

DR. NELSON: Yes, thank you.

MR. BALLOU: So the answer is we're looking at a group-level measure but to the extent that it impacts parties, for example, potentially going forward with respect to payment it's the physicians who are initially affected. So you have --

DR. NEEDLEMAN: Primary care

clinicians that are affected.

MR. BALLOU: It's all physicians in the group.

DR. NEEDLEMAN: It's all primary care clinicians in the group.

MR. BALLOU: It's all physicians in the group. There's a group score so it's a group measure in that respect.

DR. NEEDLEMAN: And you are counting the services provided by physicians assistants and nurse practitioners?

MR. BALLOU: That's correct, as well as --

DR. NEEDLEMAN: Who are increasingly operating as independent practitioners?

MR. BALLOU: That's correct. We're counting the services --

DR. NEEDLEMAN: And their services, their patients are being counted within this. It is primary care clinicians.

MR. BALLOU: Well, the patients

are attributed to the group. I may be misunderstanding the question. And the group is given a score based on all of the services used by the patients given the group which include all services, including those not even provided by the group.

DR. NELSON: To clarify, Dartmouth was part of the CMS PGP program so -- and we're an ACO pioneer so we've lived with this attribution method for about 5 years now.

It is true as you said we have about 800 physicians in our group, approximately 300 primary care physicians. We have about 35,000 attributed patients. And it's all of the group practice that's responsible for all of those patients even though they come into the attribution having had a primary care service.

Some other questions and comments. Let's see. Your card was up next.

DR. MCHUGH: I think you've answered my question. That's fine.

DR. NELSON: David.

DR. GIFFORD: So I'm just a dumb surgeon but it makes me feel good because I'm going on the same thing that Jack is going on.

So I think that Gene's example of a large group which has multi-specialties in it makes a lot of sense. That's an ACO. But I wonder what happens to, first of all, providers who are in a small group. So if I'm a primary care doc, I'm in a group of five guys, does that mean it doesn't apply to that group at all and so those providers are out of the game?

And then my second question is we have -- there are a fair number of large single-specialty groups. So let's say you're a large single-specialty group and you're referring your patient out to specialists. But those costs are going to be attributed to you, is that correct?

MR. BALLOU: I guess I'd like to give this one to Dr. Rich and perhaps after

that CMS can comment on maybe any future plans for expanding the measure.

DR. RICH: Yes, so the first question. Yes, the attribution method is intended to provide information on per capita costs of beneficiaries to that medical group that looked through the attribution method to be providing the primary care role for that beneficiary.

And like any per capita cost measure there could be a single-specialty family medicine group that is not providing all of the services. It's a single-specialty family medicine group. But now these primary care clinicians are for the first time given the information about the per capita costs of all the patients that at least according to this rule appear to be their primary care patients in the Medicare fee-for-service program.

It's also true that you might be a single-specialty cardiology group that you are

also being attributed Medicare beneficiaries, recognize that you would only be attributed Medicare beneficiaries if you -- if those beneficiaries had no visits with a primary care physician that year. Because those beneficiaries would have been attributed to a primary care group.

So the idea is then, well, you're a single-specialty cardiology group. We're all aware that there are some medical specialists and surgical specialists who fulfill the primary care role for their patients. And if there was no primary care physician involved in that beneficiary's care, then if you provided the plurality of these primary care type evaluation and management services that single-specialty group gets attributed.

And then this -- the issue that was raised earlier regarding -- regarding non-physician clinicians. It's my understanding, and Sheila may want to comment, that the

legislative requirement indicated that basically it needed to be -- this was intended to be a physician program and that's why the first step attributes patients to physicians.

But in the second step if there were no primary care physicians involved then the -- then PAs or nurse practitioners who provided the plurality of primary care services could be the basis on which the patients were attributed.

DR. PENSON: But the level of reporting then is still the group and not the individual physician.

DR. RICH: But the attribution is based on individual physicians. But the reporting is to the group.

DR. PENSON: But it reports at the group level. Gotcha.

DR. NELSON: And we're trying to work on this first criteria but I think there's some for many of us trying to figure out how this measure basically works as well

as this first criteria. So we'll be pursuing both of these for a bit. Bill.

DR. WEINTRAUB: So I think you sort of can't do one without the other without so having some sense of how this is going to be done to get a sense of why it's important.

So David, my knuckles are dragging lower than yours as a cardiologist and so I'm going to make a little bit of trouble here and say I'm not convinced that this measure makes any sense whatsoever.

You're going to have groups of different sizes and different compositions, some of which is primary care providers but some of which are multi-specialty groups and some of which are -- just have cardiologists. So they're just specialists.

And they're going to be -- they're going to attribute to that group a variety of types of services of which they may have nothing to do with it whatsoever. So that the patient who's seeing primary care for the

colds, that patient then goes and has a myocardial infarction, is hospitalized 100 miles away and taken care of, completely different sets of doctors and has angioplasty. It's complicated and they have coronary surgery. And then they go to a SNF for 6 months. And all of this is being attributed back to the primary care group.

So I'm sorry to be difficult but I'm not convinced that this measure gets at anything that's really going to be helpful to us as a society in understanding and controlling healthcare costs and improving outcomes.

DR. NELSON: Andrew.

DR. RYAN: Just to change gears a little bit. I know there's been various concerns about using tax IDs to identify groups. And I remember on the CMS GEM project there were some, you know, they went ahead and did it with some caveats that this might not be the best way to identify groups. So I was

wondering if the measure developers could present any evidence that they think the tax IDs are accurately reflecting the physician practices.

MR. BALLOU: We are aware of the concern. We think that it is a legitimate concern. Where we are starting right now we have the tax ID number available to us as something that we can use to characterize groups. In some cases that's going to be accurate, in others most likely not.

DR. WALKER: Andrew, could you explain what the concern is?

DR. RYAN: Well, I think there's potentially false positives and false negatives. I think there's some idea that some small, smaller physicians that aren't really a group would bill under the same tax ID. This is more of the concern I've heard so that some solo and small physicians could look like they're part of a bigger group when they're not really.

MR. BALLOU: The only thing I would add to this is that the tax ID is ultimately used as a unit of payment and ultimately where we're going towards is rewarding groups for high-value performance. So it will remain relevant there. But the concern is a legitimate one that you express.

DR. GIFFORD: While I think Bill and I may agree on some things I'll disagree on this one part. I think conceptually this is an important measure. And I think do physician groups need to understand where patients are going? We sort of have always sort of let patients come in to us and say when they come into us we'll take care of them, not proactively, how do we address it and where do we coordinate the care that's going out there. So I think conceptually this is an important measure. It's trying to measure an important issue.

I have all sorts of questions about the attribution. I have all sorts of

questions about how it's calculated and whether the measure results are meaningful. But the concept that it's trying to measure, what it's trying to do it passes to me very high on everything else. It's really the devil's in the details.

And if you put those together I had trouble with the importance. If you separate them then it's easier to say that there's a higher importance here.

DR. NELSON: Let's go to David and then Daniel.

DR. REDFEARN: This is a question for the developers. Is there not a rule that says you only assign and use groups that have 25 physicians in the group and at least 20 patients assigned?

MR. BALLOU: Twenty-five eligible professionals at least one of which must be a physician.

DR. REDFEARN: Okay. That would mean to me that this is probably really a

group even if you're driving it off tax IDs. Thanks.

MR. WOLFSON: This seems to be a measure that follows legislation. Could we get some background on the legislation? I mean the 25 is not an artifact of anything but legislation I believe. So I think it would be important for us to hear that because I think this is fulfilling a legislative need that we should understand.

I also want to say that this is a wonderful opportunity for PCPs, primary care, to step up to the plate. So I don't buy that they shouldn't be responsible for the care that's received. This is all about coordination of care and primary care is supposed to be doing that. So I don't buy that responsibility model.

DR. NELSON: Would you like to respond to Daniel's first question?

DR. ROMAN: Yes. The legislative history actually goes back to 2009 and

predates the Affordable Care Act. And includes legislation for -- to provide physician feedback -- that legislation included a mandate to provide physician feedback reports that gave information on cost. And quality was actually secondary in that initial legislation.

With the Affordable Care Act there were then two mandates. One continued essentially the physician feedback reports and the other mandate -- and was very clear about that being relevant for both cost and for quality. And the second mandate was for the value-based payment modifier.

MR. WOLFSON: -- specified the 25.

DR. ROMAN: No.

MR. WOLFSON: So how did that come to be, the 25? Was it a methodological cutoff or was it?

MR. BALLOU: This was specified in the rulemaking process. In terms of how it came to be we were engaging in reliability

testing. There's the question of whether this attribution rule, for example, even if appropriate for larger groups is appropriate for smaller groups, et cetera. So there's a combination of conceptual issues that we were thinking about that seemed to make more sense for larger groups. And also the reliability of the measure as well.

DR. NELSON: In the application materials I saw terms like "optional," "confidential," "phased in." And starting with groups of 200 going to groups of 25. And so I think there's a trajectory and a process of phased implementation behind this as well that may be helpful for everyone to understand.

DR. ROMAN: Yes, that's true. The physician feedback reports have been confidential reports to date to the physicians and remain so at this time.

The phase-in that I think is being referred to is for the value-based payment

modifier. The initial application of the value-based payment modifier to the physician fee schedule will begin in 2015. We have stated in our -- and finalized in our rules that it will initially apply to groups that have 100 or more eligible professionals. And that will be applied in 2015 based on 2013 performance year.

But it will phase in between 2015 and 2017 to all physicians. So that by 2017 all physicians will have the value-based payment modifier applied to their physician fee schedule reimbursements.

MS. TIGHE: Sorry, I'm just going to jump in. I know we spent so much time at the micro level on the last measure. This criteria is really looking at the measure focus. The conversations that we're having now really are more relevant to the next criteria, the scientific acceptability.

So, I do just want to pull everyone back, save these conversations and

refocus on what we're discussing here. And once we get through these votes we can revisit just to echo David's line separating the importance from the technical specifications. Let's start there and then move into the technical specs.

DR. ROMAN: I'll just address this last question. It's physicians as defined by Medicare.

DR. NELSON: PQRS.

DR. ROMAN: Physicians which is -- physicians including MDs, ODs, doctors of dental surgery and a couple of others. But not the other types of eligible professionals. That may occur after 2017. It depends on legislation.

DR. NELSON: Other questions or comments related to impact, improvement and intent? Yes.

DR. GARRETT: So my comment is about outpatient pharmacy costs, that they're not included. I understand the practical

reasons for that but it really is a limitation when you're trying to measure total cost of care. They have a huge impact on health outcomes and ability to really improve health in a population. You know, somewhere around 20 percent of cost that we're missing here. So something to think about.

DR. NELSON: And Brent.

DR. ASPLIN: This is stepping into MAP a little bit. Just so I understand the intent of how it would be used though.

If you have a 110 eligible professional single-specialty group that bills under its tax ID number and say it's a cardiology group. And a small portion of their patients attributes to them because they don't have any primary care visits and they're exclusively managed by their cardiologist.

As the value-based payment modifier is developed for the other specialties which measure would trump as the denominator and measure resource use or

expenditures? The ones that are developed on registries for the specialties, or this all -- total per capita cost measure?

And I know -- we can handle it offline if you want, I just -- it helps me understand. Because if it's this one, I mean I would hope that for single specialties we probably would have more focused measures of resource use for the denominator portion of the value-based modifier by 2017 than this. When you think of all the activity that that cardiology group would do and what a small portion of the resource use this measure would address.

DR. NELSON: David, is your card up?

DR. ROMAN: Currently for resource use specifically we have this measure and total per capita measures for chronic diseases one of which is CAD that would be included in the composite for cost in the value-based payment modifier.

DR. NELSON: Jennifer?

MS. EAMES-HUFF: So the intent of the measure is to assess physicians that are in a particular -- at least meet a minimum threshold for a medical group size. Can you talk about how many physicians do not fall under that? So how many does this apply to and how many does this not apply to?

MR. BALLOU: I can only give rough numbers right now. So, for example, the reports that we are getting ready to put out, there are about 7,000 groups nationwide that have 25 or more eligible professionals. And those groups together capture about three-quarters of all physicians billing under Medicare. So it's a minority of groups as defined by taxpayer identification numbers but of course it's the biggest ones.

DR. NELSON: Okay. Any other comments on this? Dan, your card's up. Are you good? Okay. So we've been discussing impact, improvement and intent. Further

comments before we have a vote?

MR. WOLFSON: Just a note about how many doctors practice in groups of less than five. I think it's around 60-70 percent. So let's face it, this is to reward group practices with a value modifier. And that's why we have the 25. We're not trying to take in all the physicians in the country. We're only trying to adjust for those ones that are in group practices above 25. Or we wouldn't be measuring this like this. We would be measuring much less. That's where all the doctors are.

And that hasn't changed. I mean you think that group practices is growing but that number of disproportionate people in groups of less than five has been pretty stable. And ACP could give you that kind of information. I mean we are not trying to get the universe. This is a group practice score.

DR. GARRETT: But you're required to do that by regulation, right? This measure

has no leeway because regulation requires it to be 25. Did I understand that correctly?

DR. ROMAN: The legislation does not require it to be 25.

DR. GARRETT: The regulation did though.

MR. BECKER: I thought I heard you say that you're starting at 25 and by `17 you're going to start to move it down. So you're piloting a group and getting some experience and then over a period of a few years you'll move it to the rest of the folks.

DR. ROMAN: Yes. Actually we're starting with groups of 100 or more eligible professionals for 2015 and by legislation we are to include all physicians by 2017. All physicians.

DR. NELSON: Another aspect of this topic has to do with disparities and the use for disparities, understanding differences. Any comments from the committee on that? Or questions?

DR. WALKER: That was my homework as well. So the developer said they examined per capita cost by various demographic characteristics and found no consistent pattern. And I would ask them to elaborate on that a little bit.

MR. BALLOU: We tested it a variety of different ways. We looked in particular at breakdown by race, in particular white versus non-white, and dual eligible status, dual eligible versus not. And we ranked groups in terms of the proportion of their assigned beneficiaries. For example, proportion of their assigned beneficiaries who were duals.

And then we did testing in a number of different ways. I believe what you're looking at in the report are deciles. So we looked at who was in the mean per capita costs among those in the bottom decile meaning having the fewest dual eligibles attributed to their group versus those who were in the top

decile, meaning those who had the most attributed to their group. And we found, the numbers are in here, relatively modest in our view differences in the 9,000's versus the 10,000's on a risk-adjusted per capita basis. And then we did the same thing for white versus non-white and similarly found relatively little difference between the bottom and the top deciles at the mean.

DR. NELSON: Any further questions or comments?

DR. ASPLIN: I'm just reading body language but I don't think a lot of us believe that. We believe what you're saying, it just doesn't make sense. Help us understand I guess. I believe what you're saying. I'm trying to understand it.

MS. WILBON: Can you use your microphone, please?

MR. BALLOU: To answer the question. So how could these numbers be essentially I think is the question. We

haven't done a thorough examination breaking down the numbers but by and large you would expect these numbers to be larger if the same groups tended to get all of the duals and the same groups tended to get none of the duals. And that's certainly going to be true for some groups. But to the extent that there isn't a lot of variation, you know, at least across these deciles perhaps these numbers are this compressed.

The other thing is that we're comparing -- I'm trying to give you a better answer than the one I just gave. The other thing is that we're comparing after risk adjustment here. And so risk adjustment for this particular model does include dual status. So it's possible that if we had looked at observed as opposed to risk-adjusted costs we might have seen a larger disparity.

DR. NELSON: Any further questions or comments before we move to votes? Okay. As we did this morning we have votes on 1a,

1b, 1c and then an overall vote. And we would ask for a public comment before the overall vote.

So the first issue then is impact, high-priority impact. And the numbers affected by the measure. Does the measure demonstrate variation in resource use or overall poor performance? Are there patient societal consequences of high or low resource use? So those are what we'd like to keep in mind for the first criteria. Ready to vote?

MR. WILLIAMSON: We will now vote on criteria 1a which is high-priority. You will have 60 seconds. Begin now. And we have 20 high, 2 moderate, 2 low and zero insufficient.

DR. NELSON: Okay, thank you. The next criteria, opportunity for improvement. Do the data demonstrate a distribution of performance scores? Is the number representative of the entities? Is there data showing disparities in the use of resources or

cost of care? What's the size of the population at risk? Potential consequences.

MR. WILLIAMSON: We will now vote on criteria 1b, the opportunity for improvement. You will have 60 seconds. Begin now. Eleven high, ten moderate, four low and zero insufficient.

DR. NELSON: Thanks, Evan. And the next criteria is -- scroll that up please to 1c. Oh, 2c. Do we go to 2c or 1c? 1c, okay. Okay. So the intent of the resource use measure and the measure construct, are they clearly described? The resource use categories, types of resources and costs that are included in the resource use measure, are they consistent and representative with the intent of the measure?

MR. WILLIAMSON: We will now vote on criteria 1c, the measure intent. You will have 60 seconds. Begin now. Eight high, thirteen moderate, four low and zero insufficient.

DR. NELSON: Okay. Thank you for those votes. Let's hear if there are any public comments from the people in the room or on the telephone.

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

DR. NELSON: Okay. So then we'd like to have an overall vote on this -- right, on the domain importance to measure. And you have the usual choices, 1, 2, 3, or insufficient.

MR. WILLIAMSON: We will now vote on the overall importance to measure and report. You will have 60 seconds. Begin now. Eleven high, ten moderate, four low and zero insufficient.

DR. NELSON: Okay, very good. Thank you. So we've moved through the first topic. And next we go to reliability and testing, and the whole area of scientific acceptability.

So we'll begin with reliability specifications. And we'd like to have Brent and Mary Ann lead off the discussion around topics of construction logic, clinical logic and adjustment for comparability. So who'd like to go first? Brent? Mary Ann?

DR. ASPLIN: I can go ahead. It's -- overall I'll just start with the ratings. The vast majority of committee members ranked the construction logic, clinical logic and these adjustment categories as either high or moderate in all of the criteria.

The reliability testing which may be the next phase of this was rated a little bit lower. And in general I would say that some of the comments, the written comments as well kind of mirrored the first measure in that they probably raised some more validity questions than they do reliability questions.

So I don't personally have significant questions although I do believe that based on the voting and some of the

written comments, and I don't know if this is 2a1 or 2a2 but perhaps some additional comments from the developer around the reliability testing and the results of that would be helpful for the committee.

DR. NELSON: Mary Ann?

MS. CLARK: Sure. Additional comments. I mean I guess I felt like this particular measure, the way -- the construction logic is very confusing. I mean I'm glad that you are here to help explain it a little further. But the way it's explained in the documentation I don't know that the typical user would be able to follow it very easily.

It's still -- I would like to see a little bit better explanation of how it's put together, especially for the attribution. That's primarily what I'm talking about. So in certain places it talked about primary care services but then it also said other. Well, actually it was talking about the requirement

of having at least two office or other visits, other outpatient services. And then there wasn't really a definition of what that meant. So I know there was a listing of different E&M services there, but is that the inclusive list or were there others that were used to define this particular criteria?

And also I was a little confused on the data being used to create this measure. It looked like you were using -- was it just a test case? Because it looked like it was only nine states or something like that worth of claims data. So now it's being expanded out, is that right? Okay.

I am still a little concerned about the inclusion -- mainly the exclusion criteria. And again this is one where patients who died are excluded. Patients with outlier costs are excluded. So it's kind of the same concerns that I had on the first measure as well.

And you know, it seems like those

are something that should be considered because those are the areas where there may be the most room for improvement. So I think those are my comments.

DR. NELSON: Thanks, Mary. Replies from the developers?

MR. BALLOU: Sure, I can give some initial replies and others might join in.

First of all, the reference to two or more office visits unfortunately is a typo that shouldn't have been in here. So the codes that are listed in the appendix are the correct codes. It is the two-step rule that we started off the conversation by talking about.

In terms of the exclusions we have in previous iterations included what we've referred to as part-year beneficiaries. And that's largely what we're talking about with the exclusions, not exclusively. And those are the folks for whom we only have part of a year of either Part A costs, Part B costs or

expenditures or both.

The way we've considered doing it in the past is we've done weighting in our risk adjustment model that takes into account the number of months of data for which we do have them. But there's an implicit imputation there when we try and include these part-year folks that was not well received by physician stakeholders.

That is to say you're telling me that you're judging me on a full year of care when I only provided 6 months of care to this person. And you're simply doubling that. And you can do more sophisticated imputation than that. But the imputation in general is something that was, again, not well received.

And so I think largely for that reason the part-years were excluded. We realize that there are some implications, however, which the commenter was just mentioning to doing so.

DR. NELSON: Jack?

DR. NEEDLEMAN: I need a clarification of how the measure is constructed. In the previous measure that we were talking about there are a whole series of indicator variables for the different HCCs, for the age categories, for the DRG and so forth. There's a regression of the standardized billings on those indicator variables and that produces the ability to get an expected value.

I thought that was what was going on here but then I see some reference to an HCC score and one part of your documentation has a score which looks nothing like an expected dollar value. It's in the orders of ones and twos.

So can you explain how the risk adjustment score is actually being constructed in this? Because it does not look like an expected dollar amount based upon a regression on indicator variables.

MR. BALLOU: No, that's correct,

it's not. And essentially the HCC risk score is an output of a series of CMS HCC risk-adjusted models. We do not construct that risk score. It is a value of -- this is a high-level summary -- of 1, less than 1 or greater than 1 indicating expected costs for that beneficiary.

We take that risk score -- so for example, a risk score of 1.5 would be a beneficiary who based on prior year diagnoses and demographic information would be expected to have costs 50 percent higher than the mean.

We then take that risk score and we also take its squared value. And this is meant to get at some outlier issues a little bit, and we also take an indicator for ESRD status. And we regress our standardized billings on those variables. So that's a high-level summary.

In terms of how the risk scores come about that would be a question for Greg Pope on the phone about the CMS HCC risk

adjustment model. But that may not be what you're asking.

DR. NEEDLEMAN: No, no, that's exactly what I'm asking. Because we've got this composite score and I don't know where it came from. So Greg?

MS. TIGHE: Operator, can you make sure Greg Pope's line is open?

MR. POPE: Hello, can you hear me? This is Greg Pope.

MS. TIGHE: We can.

MR. POPE: Okay. So the CMS HCC score is calibrated on Medicare fee-for-service data regressing a beneficiary's current year expenditures, his or her prior year demographics and diagnostic indicators. And then that does produce a predicted dollar expenditure but then it's divided by the sample mean to turn it into a relative score with 1.0 indicating an average predicted expenditure, 1.5, 50 percent above average and 0.05, 50 percent of average.

DR. NEEDLEMAN: Okay, so when we see the full risk model described in the documentation with the HCCs, the age categories interacted with sex and the others basically the expenditures for that individual is being regressed on indicator variables for all those, producing an expected value and then that's divided by the average expected value for the whole sample to get that ratio --

MR. POPE: Correct.

DR. NEEDLEMAN: -- of expected for the patient -- and that's what's going into the second stage risk adjustment model.

MR. POPE: Correct. That's correct.

DR. NEEDLEMAN: Okay, got it.

DR. NELSON: Other questions or comments? David.

DR. GIFFORD: Can I ask about the attribution here or is that a different section?

MR. AMIN: If it's a question around the preciseness of the attribution this would be a place to discuss it, or if it's a question of the validity or the appropriateness of the attribution that would be a validity question.

DR. GIFFORD: I guess I don't understand the attribution question to be able to discern that. I just need to understand how the attribution works and then maybe I may or may not have questions on it.

So you used the E&M codes and the charges to attribute. And the maximum amount got you into the group practice, assigned to a group practice. Is that my understanding?

So my interest is this understanding of all internists are not primary care physicians. And I guess that's where I'm looking at it. So maybe it goes under the validity question. So I can wait till then. Sorry, I had to talk that through in my head.

MR. AMIN: I don't mean to be overly restrictive but if we can.

DR. NELSON: More on reliability. I think Herbert was going to open up the topic of reliability testing.

DR. WONG: So in terms of reliability testing overall the pre-testing indicated that folks were basically comfortable with the approach. It was pretty much divided between high and moderate, 9 with high and 11 with moderate and 1 person voting insufficient.

The way I would characterize it is that for the two comments that I kind of highlighted in terms of the reliability testing one basically questioned the notion of using prior year data as opposed to concurrent data. And I think that that was addressed a little bit earlier.

I think that in general the viewpoint of those who had some concerns with reliability was this notion that it was pretty

much the standard sort of package and they wanted to kind of see more in terms of some of these other components to kind of highlight the fact that the reliability is really on target. I would characterize it in general that folks were relatively comfortable with it.

DR. NELSON: Okay, thanks Herbert for that opening. Questions, comments on reliability testing. Bill.

DR. WEINTRAUB: Well, this sort of gets back to what David was talking about also about reliability and validity. And so it's worth remembering that you can be very precise and totally inaccurate. So this may seem to work and be reliable but we may not be measuring what we think we are. And we'll discuss that in the coming section.

MS. WILBON: Gene, just real quickly. Just so everyone's on the same page about what the reliability testing was I'll just ask Carlos to kind of summarize what he

found in the testing and kind of give an overall assessment just so everyone is on the same page.

MR. ALZOLA: Okay. Thank you, Ashlie. My comment on the reliability was that we have two components here. We are looking at the size of the groups in terms of EPs and then we have the size of the groups in terms of beneficiaries.

So the reliability testing that was done was based on reliability scores. And again the -- as in the previous measure when we look at the individual groups the question was how high was the reliability score in each of those cases.

There were -- the analysis was presented in terms of the size of the groups. And we have, for instance, we have the category of 25 to 50 EPs, and then we have for that group we have that 99 percent of all groups exceeded the 0.5 reliability score. And for -- and 93.2 percent of all the groups

exceeded reliability score of 0.7.

So in general all those numbers held across the size of the group. As the groups grow larger the reliability scores improved. For the very large groups with 201 or more EPs the reliability was 100 percent. I mean, all groups had reliability exceeding 70 percent.

The other way of looking at that was when we look at the number of attributed beneficiaries on the group. So they look at the lowest quartile from 20 to 249 beneficiaries and then we have that in that case 97.5 of groups had reliability exceeding 50 -- 0.5. And 83.6 of groups had reliability exceeding 0.7.

As the number of attributed beneficiaries increased we move onto the second, third and highest quartile, then all groups had reliability exceeding 70 percent. So based on those numbers and that criteria we can say that the reliability is very high.

Any questions, comments?

DR. NELSON: Thank you, Carlos. Any questions? David.

DR. GIFFORD: Just reliability about the attribution. And I know that patients are changing but how well did that stay over time, repeat over time? And did group size being defined as 25 change over time too?

MR. BALLOU: So we did not test actual changes over time. We had the one current year of data and we used the signal to noise approach to do it. Changes over time have not yet been tested.

DR. NELSON: Other questions or comments on reliability? Okay. Okay, any public comments? Nancy.

DR. GARRETT: I have a question about risk adjustment. Would this be the right time or is that the next?

DR. NELSON: We heard it depends on what you ask.

(Laughter)

DR. GARRETT: All right, well I'll defer then. I'll ask in the next section.

DR. NELSON: Nancy, go ahead.

DR. GARRETT: So, as I understand it in this model it's adjusting for sex and for dual eligibility status. And I think it's perhaps because that's how the CMS HCC model works. But for us who have been here all day in the last discussion we heard that we weren't adjusting for those things because it was against the conceptual idea of -- the NQF position on risk adjustment where you try not to adjust for socioeconomic status up front which I don't know if all of us agree with that. But it sounds like that was the -- kind of some of the thinking. So I just wanted to hear from the measure developers about the philosophy behind this.

MR. BALLOU: Right. Greg, if you are still on the line would you be able to address the question of the inclusion of dual

status in the model as well as age and gender?

DR. GARRETT: Specifically gender. I think age is a more clear one.

DR. NELSON: And Greg, are you on the phone?

MR. POPE: Yes. Can you hear me?

DR. NELSON: Very good, thank you.

MR. POPE: Yes, just from our point of view I think as has been mentioned this model was developed for Medicare Advantage and those were the historical demographic factors that have been used by the CMS actuaries going back to the old adjusted average per capita cost.

So it's carried over and it's been found that those factors predict cost even aside from holding constant diagnosis. So they've been carried along and I guess the age/sex in particular are regarded as exogenous or not subject to manipulation by health plans. So you know, in the Medicare Advantage context we want to avoid selection

incentives for plans to avoid certain types of beneficiaries based on their age or sex. So that's the rationale for including it in that context.

DR. GARRETT: And dual eligibility status?

MR. POPE: I think it's similar, that that's been found to correlate with cost even holding other factors constant. And it is regarded as sort of a low SES type indicator to avoid giving plans incentives for avoiding enrolling those people if you didn't adjust for that. We don't want plans to want to avoid dual eligibles.

DR. GARRETT: So what I'm hearing is those variables are in the model really because it was developed for the Medicare Advantage purpose, not for the purpose of this measure. But it's being used in this measure. Is that right?

MR. BALLOU: It is being used in this measure. I would add to the information

that Greg has already provided that the ACA actually requires us to adjust for SES characteristics. And we view the inclusion of this variable in the model as one way of accomplishing that.

DR. GARRETT: Interesting. Okay.

MR. BALLOU: Sheila, did you want to add to that?

DR. ROMAN: Basically I was going to say the same thing, that the ACA asks us to adjust for SES and for demographic characteristics.

DR. NELSON: Thank you. Carlos?

MR. ALZOLA: I have a question. I have heard that the SES and dual eligibility is in the model. But when I look at this documentation it clearly is not. So I'd like to clarify whether it is or not.

MR. BALLOU: Yes, I regret the confusion this might have caused. It is in there in the sense that it's included in the risk scores which are then inputs into the

second-stage regression. So that risk score is informed by, among other variables, the dual eligible status.

MR. AMIN: I'd also like to just clarify in terms of the NQF criteria which has been brought up a number of times now which is in the validity section but for the sake that this has been discussed a few times. I just want to be crystal clear in terms of what NQF's position is related to these factors.

So when we look at the next criteria under validity which is 2b4 we require that any risk adjustment strategy is evidence-based and that they are based on patient clinical factors that are associated with -- that are present at the start of care to avoid this issue around the fact that there are components, there are factors that could occur during the measurement period that we don't want to create the circular issue which is that then they're included in the risk adjustment model and the overall predicted

value would go up.

Secondly, related specifically to the issues around SES, race, or potentially dual eligibility status, the concern here is that even if there's a recognition that it's related to the outcome of interest in the application that we're looking at right now which is a performance measure we want to make sure that there is an understanding of -- that there's no obscuring of the disparities of care, that they should be highlighted, either stratified and not buried in terms of the risk adjustment model. Because if it is in the risk adjustment model you're not able to actually see the disparities across the different population groups.

So the question that's in front of the committee to think about as you're thinking about the risk adjustment model is whether or not that -- these variables, particularly dual eligibility status in this particular application is appropriate given

that if there is actual variation on dual eligibility status that are outside of patient clinical factors is there an a priori reason -- or is there an appropriate a priori reason why a provider should be having a difference in resource use based on the fact that you're dual eligible outside of patient clinical factors. That's the critical question that's in front of us.

Generally it has been NQF tradition and guidance to avoid having these characteristics in the risk adjustment model for the reasons of obscuring disparities. Now, if there is a justifiable reason in this space why that would be necessary we need to have that conversation in the context of the validity of the measure.

And that is a serious question to be considered by this committee. That would be different than the tradition of the past but if it's considered to be appropriate in this context we can have that discussion. But

thank you for bringing these up, it's a very important question. Hopefully that was clear.

MS. TIGHE: And I think actually that said if we could hold the validity conversation and get through the vote on reliability.

DR. NELSON: Public comment? Any public comments?

OPERATOR: At this time if you would like to ask a question please press *1 on your telephone keypad. There are no questions at this time.

DR. NELSON: Thank you. So, we've heard discussions about -- under the topic of reliability about exclusion criteria and methods of construction and attribution. And the specifics of the reliability testing methods that were used and this issue of dual eligibles being in the adjustment model.

So I think we're ready to vote. And this vote, it's all things considered for reliability what's your assessment, high,

moderate, low, insufficient. So I think we're ready to vote.

MR. WILLIAMSON: At this time we will vote on overall reliability. You will have 60 seconds beginning now. Five high, eighteen moderate, one low and zero insufficient.

DR. NELSON: Okay, thank you. So the second half of scientific acceptability has to do with validity testing. And we have several topics here that we'll be addressing. The specifications, construction logic, clinical logic, adjustment of comparability. And then validity testing exclusions and risk adjustment, identification of statistically significant differences and substantial substantive differences and disparities. We opened up the disparities conversation earlier. We can return to that as helpful.

So to open the discussion on validity testing, Jim.

MS. TIGHE: Sorry, and Jim, we'll

start with the validity of the specifications and then.

DR. NELSON: Sorry, I dropped down a row. So validity of specifications. Thank you, Lindsey. So Matthew and Lisa.

DR. MCHUGH: So Lisa and I split our work. I'm going to start with the construction logic. We reviewed the seven-step logic. I'll say that the overall rating for this was 16 high or moderate and 4 low.

The concerns that were raised some of which we have gone over, I'll just kind of summarize a few of these around attribution logic. Again we've covered a number of these but there were questions about whether a more appropriate approach would be to assign cost proportionally. There are questions about the match with intent in that the measure only covers groups with 25 eligible professionals and 20 attributable patients.

A number of concerns raised around the use of services provided by nurse

practitioners, physicians assistants being only accounted for in the second stage which may increasingly be outmoded as primary care changes. So that's maybe excluding a large proportion -- a growing proportion of practices.

There may be issues with patients that leave certain practices, particularly if they differ in terms of their cost profiles. And there was also concern raised that the measure does not capture a number of important costs, potentially drug costs and unbilled services around patient care coordination, education, other kinds of drug -- well I mentioned drug expenses and in kind services.

DR. LATTS: I'll take on the clinical logic and adjustments for comparability inclusion and exclusion criteria. As far as clinical logic goes there essentially isn't any. It's not really directly relevant to this. It's just total measurement for all situations for the year.

In terms of inclusion/exclusion criteria it is very similar to the other measure this morning in that it is a continuous enrollment for everybody in Part A and Part B for the entire calendar year. So if you are out for any reason including death for part of the calendar year you're out. If you were excluded earlier from attribution and we've talked about some of that you're not in. If you are Medicare Advantage for any part of it it's in. It is Part A and B as I said so it does not include any Part D claims. If you're not in the United States you're not in. So those are the main inclusion/exclusion criteria.

The voting on both of those were quite high, 18 for the high or medium for both and 3 for the low.

DR. NELSON: Thank you, Matthew and Lisa. Questions, comments on this topic? Jack.

DR. NEEDLEMAN: In contrast with

my usual calm, reasonable self I have fundamental problems with the attribution model here as it relates to nurse practitioners and physician assistants. They are increasingly providing primary care services. They are increasingly for a large, significant portion of the population, maybe not a large portion of the Medicare population yet their primary care providers. They belong in the first stage of this model, not the second stage.

I can imagine a case somewhat extreme but we can -- of a community health center, heavy nurse practitioner, heavy physician assistant doing primary care, an occasional consult out of the group for consultation on a diabetes care or an arthritis patient and they come back after the consult. But because the only counts of primary care services are the physicians that patient gets attributed to the group that the referral for the consult was when that's not

the group delivering the primary care.

I think it's perhaps not a large problem now. As we grow the nurse practitioner and physician assistant populations, as they do more and more primary care it's going to be a growing problem. I'd rather see this measure be adapted to the future we're facing, not to the past we've been in. And frankly until the attribution is changed I will vote no on this issue. This will cause me to vote no on this measure.

DR. NELSON: As a point of information do you have a sense of what proportion of primary care is currently being provided by nurse practitioners or PAs?

DR. NEEDLEMAN: Well, from the American Nursing Association I got told that there are nurse practitioners, not the PAs, just nurse practitioners are providing services to 6 million Medicare fee-for-service patients. That's roughly 1 in 5 or 1 in 6.

DR. LATTS: But the other relevant

question here has to be what percentage of those nurse practitioners are billing independently versus billing under another physician. And that's what we need to know.

DR. NEEDLEMAN: I think this came from billing independently.

DR. LATTS: Okay, yes.

DR. RYAN: I'll just add to that. Most of these data that I'm familiar with as well are related to data. So there are enough to generate measurable, sizable numbers around --

DR. NELSON: Lisa?

DR. LATTS: Well, so I agree with Jack's comment and I think this is critically important. So I just wanted to ask the developers is this something you can do? I mean why didn't you do it up front and could -- you know, if this committee said we absolutely need to add other professionals to clinician -- to the physicians, is that doable?

MR. BALLOU: I'd like to defer to CMS on what might be doable in the future or acceptable.

DR. ROMAN: I think it's difficult for me to say with any specificity since it would have to go through rulemaking. I think you see the attribution model as it is because this was a measure that was aimed to look at per capita costs for physicians. So the physicians were the entry point into the attribution model.

I think that when we're talking about attribution we're talking about attribution to the group that does include the nurse practitioners and other eligible professionals. And that in the second stage of the attribution model we certainly are counting those eligible professional primary care services as the plurality for the assignment.

DR. NEEDLEMAN: Only if after you've counted physician services and you

haven't made an attribution based upon services provided by a physician. I'm saying that's a fundamentally flawed model, that the primary care services provided by nurse practitioners and physician assistants should be included in that first stage count of where are the plurality of services being provided.

DR. RYAN: I agree with Jack on that. I think that the -- and just by looking at claims where Medicare -- we know that they're Medicare patients. And as Jack pointed out it does tend to be more dual eligible, but where patients are receiving almost all, if not all their services from these providers. I'm trying to figure out if we are in service of the statute alone or in a fundamentally sound measure.

MR. AMIN: Gene, can I just jump in here real quick? I just want to point the committee to the fact that when we're looking at validity, and this is no pro or con anyone's position, but the question is really

whether the measure specifications are consistent with the measure's intent as described under 1c which is what the developer put in 1c.

Now, the question becomes you need to assess how accurately this actually reflects the measure intent. But I just want to make sure that we're not -- you don't want to create a measure here. The goal is to really assess what's in front of you.

DR. NELSON: Thank you for that comment. Just a perspective not as a facilitator but as a perspective on the conversation. I appreciate Jack and Matthew's point on the importance of nurse practitioners and PAs in providing primary care.

Another issue is one reason that this kind of measure is going forward is the rate of increase in per capita healthcare costs, and how to get a handle on that big issue. And so that's another factor in moving and considering this kind of a measure is that

we're trying to get a handle as a country on per capita healthcare costs and measures related to that. And the majority are associated today with physicians. And so we're into this conversation and into this measurement in part driven by this very serious per capita cost issue. So I wanted to raise that as part of our thinking about the environment that we're working in.

DR. RICH: I think there was also a question regarding sort of what the data were telling us. And this is absolutely looking at the past rather than looking at the future.

We looked at other attribution models in 2010 data. And it turned out that very few nurse practitioners in our data set were attributed patients under other sort of E&M-based models.

And our anticipated explanation for this is that, as many of the clinicians here will recognize, many PAs or nurse

practitioners in many settings bill an incident to type of code so that their billing number does not show up. It shows up under the name of a physician in that practice.

Now, clearly again that's looking at the past, not at the future. It's reflecting a time when there were few practices where there were no physicians and only nurse practitioners and PAs. But there was a question about the data and that's what the data showed at that time.

DR. NELSON: Joe?

DR. STEPHANSKY: The modeling seems to assume a somewhat stable physician marketplace in a state. And we don't have that in Michigan. I mean, we're looking at mergers between hospital systems and new tax identification numbers. We're seeing big groups split into smaller ones because they want to take on different ACO contracts. And so they have a new TIN for each one of those. It is shifting.

And we have a whole lot of really, I don't want to say elderly but approaching retirement physicians where those practices may completely split up once the principals do retire. And the idea that we're going to continue to be able to look at a physician group over a year, over an extended period of time just to me doesn't seem to match.

And that's why when we go back to Brent and his question about how many Medicare beneficiaries were attributed somehow those numbers still don't seem right to me when I think about the physician marketplace. And to me that's a validity question.

DR. NELSON: Thank you. Tom?

DR. TSANG: I was actually going to ask the same exact question. And then I have a second question for the measure developers. And that piggybacks on Jack's point is that a large sector of the primary care services are provided by a few HCs. Jack mentioned some of the nurse practitioners

actually work in them. So they have a different Medicare payment system. They get paid by a Medicare PPS system. And that -- and a lot of the FQHCs actually don't even use EMM codes and CPT codes. So I'm just wondering how this measure would actually account for a large sector of the primary care services being delivered by FQHCs.

And the other thing is that FQHCs will double to roughly about 2,300 or 2,400 over the next 2 or 3 years. A large segment of the 30 million that's uninsured that's going into the exchanges will probably seek out primary care services in the FQHCs as well as the rural health clinics. So I just want to understand the context as well.

DR. NELSON: Great question.

MR. BALLOU: So in terms of the FQHCs you are right that they do bill different codes, they do bill in a sense under a different system. So we, again, when driving toward value-based modifier related

work we're talking about modifying payments to the physician fee schedule. So attribution is based on physician fee schedule services only.

It is true then that a lot of the FQHCs, they are not picked up. That's something that we are looking at, how one might incorporate going forward. But they're not there right now.

DR. NELSON: Bill?

DR. WEINTRAUB: Most of the discussion so far is on construct validity. And there are lots of questions being raised around the table on stability of the groups, what to do about nurse practitioners, what to include in the risk adjustment model. I'd actually give kudos to the development team for trying to do the impossible here because I think that the fundamental problem is not with construct validity but with face validity and criteria validity.

The face validity of this to me makes no sense as I said before. It's sort of

the interplay between the problems of validity and the problems of the importance of the question then.

In criteria validity we have no gold standard to measure this against. So I think we're really in very deep water on this measure.

DR. GIFFORD: Now I can ask my attribution question. I think it would be helpful to see, since you're saying this attribution by cost is better than by visits, and some of us are familiar with seeing the visits, see data that shows why you think it's superior, better, or different. Or what would happen if you did it a different way. I didn't see any of that there. Maybe it was on some of the website links I couldn't get to.

The other is I'm still trying to understand this, the internist question. Now I know how geriatricians and primary care internists bill because that's what I am. And I was in a large group practice with about 50

or 60 physicians and about 20 of us were primary care and the rest were specialists. And I should know how they bill, but. Are they using the same E&M codes, 211 through 215 for follow-up visits?

DR. RICH: Yes, particularly -- so the outpatient evaluation and management services, some of us will remember when there were E&M codes that were for visit codes and there were E&M codes for consult services. Those, the consult services still exist in the CPT codes but they are not reimbursed under Medicare and therefore for billing Medicare physicians providing outpatient evaluation and management services all use the same codes.

So let me speak to, I think you mentioned the internist question several times so let me unpack that briefly. I think you're alluding to the fact that there has been sometimes a policymaker assumption that an internist equals a primary care physician and you're positing that that may not be the case.

And yes, I think that this attribution model attempts to deal with that in the following way.

There can be an internal medicine physician who's identified themselves as internal medicine who's doing outpatient general medical care. So that sounds like a primary care physician.

There may be one who's a hospitalist that is not a primary care physician. There may be one who is actually working in an emergency room. That is not a primary care physician. There may be one who reported themselves as an internist but restricts their practice to cardiology or some focus area and is not attempting to do primary care for their patients.

The hospitalist, since the codes that we're looking at are only -- are not in the observation status or inpatient codes the hospitalist is out. If the physician is working in an emergency room billing emergency

room service codes they're out. So the internist working as a hospitalist and the internet working in the emergency room is not going to be attributed patients under this model.

Now, the internist who is doing mostly cardiology may be attributed patients in the same sense that a cardiologist is attributed patients except in this case that internist since they're an internist will be attributed patients for the first step of the rule rather than isolated to the second step. So if that sort of helps you sort of think about.

So the fact that the primary care services only look at outpatient evaluation and management codes, SNF, home visits, those sorts of codes. It excludes hospital visits, it excludes emergency room visits, critical care visits, et cetera.

DR. GIFFORD: Yes, so I actually assumed with the codes the ER and the

hospitalist were cut out. So my worry is, so my patient goes to -- has a pacer in, has COPD. And they see a pulmonologist and a cardiologist. They bill 211 or 215 or whatever. I see him and bill 211. The cardiologist has a few extra bills or sees him maybe more, sees it's a new patient and anything else. Their cost structure could be higher to attribute to the cardiologist than me. And you attribute it to the cardiologist group as the primary care. That's what I'm saying. How do you differentiate when you're seeing the internist?

DR. RICH: Well, so long as those internists that you were referring to designated themselves in their specialty to CMS then the first step of the rule, you would be the internist.

DR. GIFFORD: Can't a cardiologist designate as an internist and a cardiologist?

DR. RICH: Sheila?

DR. GIFFORD: I think they can,

can't they?

DR. ROMAN: I believe they can.

DR. GIFFORD: So how do you discern --

DR. ROMAN: -- primary and a secondary.

DR. GIFFORD: Yes, so how do you discern that?

MR. BALLOU: Based on the specialty that's listed on the plurality of claims. If you've listed several specialties but the carrier has placed based on the services you've provided internist on the plurality of your claims you're then an internist from CMS.

DR. GIFFORD: I would like to see, and you probably take, since you have large groups of 25 it may be worth taking a few of these at random and going in and seeing what these groups look like and how you're assigning them and doing some testing. Because I just, I just think that there's such

room for problems with the assignment.

The other question that I had with assignment is so, and this is sort of about the broader cost. My group practice of 50 that's got 10 of us that are internists, are the costs of the entire group of all 50 or just us 10, the population of our 10? How are you doing that cost for a group?

MR. BALLOU: Well again, once you are attributed a patient through the rule which again is physician triggered and it relies on relationships between individuals, patient and provider, that attribution, that patient is attributed to the group. And so from again the perspective of doing costs we're talking about then the group is not accountable for that patient's entire standardized services.

DR. GIFFORD: It's only those that are attributed to the 10 of us that you consider primary care. The other 30 who are specialists, their costs are attributed to us

but the patients they see are not included in there.

MR. BALLOU: Unless they were attributed, those patients -- or the group has attributed those patients under the second step of the rule.

DR. GIFFORD: Under the second step as the primary care. Okay.

MR. WOLFSON: It is only ambulatory care. It is no inpatient skilled nursing facility, no institutional cost. No, no, the total -- the cost is everything. It doesn't just include ambulatory, it includes inpatient. Right? Every dollar spent.

DR. NELSON: We've got about a half an hour before we adjourn. And so in order to cover more territory around this large issue of validity let's ask each person to present and then we'll come back and open it up for questions. So it would be Jim and then Andrea and David and then Herbert and Joe and then Martin and we covered Lina. So let's

get more on the table and then we'll open it up for discussion and comment again. So Jim.

DR. NAESSENS: Yes. Basically they did three assessments of validity testing, correlation of the standardized risk-adjusted scores with non-standardized payments and non-risk-adjusted totals, correlation with other utilization and a correlation with other utilization in 2010 and 2011 for a subset of states. They also did a face validity on a small sample of physicians based on interviews.

The scope of the testing for the subset of states was limited like I said to a handful of states. In that assessment there was an exclusion of a large number of practices in that only 55 percent of practices in those eight states were included.

Our evaluations of testing were actually fairly good, 3 high, 13 medium, 3 lower and 2 insufficient. Overall the comments reflected that there appears to be

some under-prediction at the top of the distribution with some high variability at the top of the distribution with the suggestion that the trimming may have been at too high of a level.

Testing correlation of the adjusted and the unadjusted values is not compelling evidence for validity. That they could have tested the correlation with other measures from other data sets. The suggestion was looking at the Dartmouth end of life data and assessing that on a practice basis.

Prediction variability is increasing as the value increases. There's more information about anticipated variation across groups -- oh, more information about an anticipated variation across groups would be helpful for assessing meaningful differences.

And there were some public comments related to validity raising issues again about attribution.

DR. NELSON: Thank you for that

summary. Carlos, do you want to add in at this point?

MR. ALZOLA: One of my comments about validity was that, one was the large number of practices that were dropped off the analysis because of the restrictions imposed.

And the other one is that geographically there is no representation of some important areas of the country. That would be the northeast and the southeast. That's the data they have.

And then I also raised an issue about the high correlation of the unadjusted --

MS. WILBON: Operator, can you mute the line that was having some breakthrough noise? Thank you.

OPERATOR: Yes, ma'am.

MR. ALZOLA: I raised an issue of I did not consider a correlation of the unadjusted and unstandardized per capita costs with the standardized risk-adjusted values and

appropriate validity tests. They were receptive to that comment.

And they presented a similar analysis summarizing all the utilization statistics and correlating them with the risk-adjusted standardized total per capita costs which show a high correlation which it was a correlation of 0.7 in 2011 and a correlation of 0.529 in 2010.

The other change they made was that in the exhibit that they presented they were correlating total costs versus the -- instead of the per capita costs which they corrected in the new table. The new correlation values ranged from a low of 0.25 to 0.705. 0.776 is the highest which is a correlation with the number of post-acute services.

In general their -- I think these values are indicative of validity. There are a couple of them in the 0.3-0.4 range but the highest one would be with the one that is the

most important which would be with all services is 0.705 which is a good number in my opinion.

So the other criteria that they used to test validity were, let me find that. They did some test of face validity. In 101 interviews with 20 of 25 physicians in Baltimore, Boston and Indianapolis. And the findings were that the physicians responded favorably to holding multiple providers responsible for the patient costs.

And in general they were in agreement with the measure. So on those grounds I consider that validity is being appropriately tested and I find it -- that's -- it has been reasonably shown that the measure has validity.

DR. NELSON: Thank you for those comments, Carlos. Andrea and David, exclusions.

DR. REDFEARN: I think we've beaten the exclusions, the horse dead, well

dead, so I don't want to go over it again. The only thing I'd mention is that Medicare Advantage patients are being excluded and anybody who resided outside the U.S. for good reasons because you're not going to capture that data.

And again indirectly through the contiguous enrollment requirement patients who have died are being excluded. We talked about that a lot for the other measure.

DR. GELZER: And I'm just going to add on the two exclusions, the newly enrolled or disenrolled, the continuous enrollment in Medicare Part A or Part B and the Medicare Advantage exclusion. So individuals have to be enrolled for that whole calendar year. As you've said that decision has been made. And Medicare Advantage individuals are excluded.

So if the intent of the measure is to measure performance, reduce variability and ultimately reward practices based on the total cost I'm a little bit concerned about validity

of the measure because of I think a potential to exclude dual eligible costs from the calculation.

I know that T testing didn't show any statistically significant differences in beneficiary demographics and duals. Dual eligible status and distribution of risk scores of those included versus excluded weren't statistically different. That surprises me like it has surprised others.

But there's sort of a double exclusion whammy because of the calendar year requirement. And then the Medicare Advantage exclusion from measurement. I think that can lead to some real gaming of the measure in that large, sophisticated practices will learn that they can actually shift some of these higher cost members to dual SNP plans and no longer have them in their -- still have them in their practice but no longer have them in the measurement.

DR. NELSON: Thank you for those

comments. And Herbert and Joe, risk adjustment.

DR. WONG: So I'll give it a start and Joe can put in his two bits on it. So I think that we've already ventured into this territory a little bit in terms of risk adjustment. From the pre-voting we had 10 high, 6 moderate and 5 low. And I think that the substance of the comments I think by going through this is this really fundamental belief that it's the standard Medicare analytics in terms of risk adjustment. And there was this notion that folks wanted to kind of see more in terms of some sophistication beyond that.

There was a general assessment that for those who are concerned that the HCC model is basically weak and that there could be some additional robustness testing to kind of explore it.

One person commented about the distribution in terms of looking at basically cost information that is heavily skewed, the

distribution could in fact be different and that different models should be considered, whether it's a negative binomial or gamma distributions, things along those lines.

I think that the regular themes that we've heard a little bit earlier with the other measure I think kind of came out here as well. In terms of inclusion of some socioeconomic characteristics or market factor characteristics because some believe that the R squared was relatively low.

So I think that that's kind of my broad summary. And Joe, if you have other comments?

DR. STEPHANSKY: Dead horse.

MR. AMIN: I have a quick question for the reviewers on this. The specific issue that was raised around dual eligibility status considering the fact that there was data demonstrated by the developers in terms of its relationship to the outcome or the lack thereof. Is there -- can you describe your

opinion on the appropriateness of the inclusion of dual eligibility status in the risk model considering the NQF guidance related to this.

DR. WONG: You know, I would let others on the committee kind of chime in on that particular aspect of it since you're honing on an issue that other folks have raised.

DR. NELSON: Brent.

DR. ASPLIN: I have a couple of other questions too. I think this issue is going to -- other than the fact that it's buried in the risk score so maybe it won't be so obvious I think it could blow up in NQF's face, frankly.

I think it's going to be very challenging because people are going to say the same thing, well, why didn't you include it then in the readmissions measure. And why didn't we put it into the risk model on the measure we talked about this morning.

Because you have the policy, and we're kind of stuck between Congress and NQF here. You know, this is an NQF meeting but if Congress mandated the sex and dual eligibility status in the risk model, although I didn't hear that about the duals. That was more of a historical reference from an earlier comment.

I just think you're going into -- you're changing and unless -- I was trying to think of a rationale for the change and was struggling a bit. But perhaps it is if you're talking about a pure quality measure you don't include those demographic and SES factors in the risk model because you want to see the disparity and you don't want to send a measure that it's okay to have poor quality if you're from a low SES status. Okay, so maybe you could distinguish quality from resource measures with the rationale on the latter being you wouldn't want to incent providers to exclude taking care of those patients.

That's the best rationale I can come up with but most people do not believe the all-cause readmissions measure is a quality measure. They think it's a resource measure, a cost measure. And so they would come right back at and want you to put it right back into it. I don't know how you square this. It's going to be challenging. Sorry, I'm not very helpful. Now you're supposed to say something brilliant.

MR. AMIN: Well, in some ways actually you're asking the precise question that I'm asking the committee. Which is that NQF guidance is very clear on this matter. And this committee if it so chooses to move forward a measure that has a marker of SES, dual eligibility by most standards is a marker of that, there would need to be a clear justification of that reason.

And this argument -- I mean some have made the argument that there's a difference in quality and resource use. The

argument being that if it's used to decide a payment as it would be in potentially Medicare Advantage and the original purpose of the way that this risk-adjusted model was created, that would be appropriate in the sense that you're describing which you don't want to incentivize providers to shirk these types of patients because you want to adjust for the cost expectation. But what is the a priori assumption here besides the patient clinical factors and making that rationale clear?

As this measure moves out of this committee into CSAC that is the exact question that the chairs and others will have to defend. And so that is what we're putting back into the discussion specifically because this is the place to have it. And the committee will have to defend that position.

DR. NELSON: There's several people that wish to comment. Let's get one more perspective on validity and then come back to these comments. Martin on

identification of statistically significant differences.

DR. MARCINIAK: Sure. So unlike the last time I spoke I'll break it out in terms of the vote to start because there was a better distribution this time. We had seven who voted high, eight moderate, four low and then two sort of in the inclusive range.

If you're looking at the remit was the data analysis sort of demonstrates the methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically clinically meaningful differences in performance.

And so we've heard through the course of the day, you know, part of this is about getting a handle on the rising sort of per capita rate of healthcare that we're currently experiencing in this country.

And so three things popped out in my mind. And frankly they're three things

we've talked about throughout the course of the day. I mean there were things such as attribution. There was the question of the measure being reported at either the group or practice level versus the individual level. And so if it's one or the other it becomes very hard to discern. And what is clinically meaningful. You know, clinically meaningful at what level? Is it at the individual? Is it at the group? How do you discern that?

Finally, with sort of -- we've touched on things that I'll link to sort of face validity and some of the challenges that we're having sort of pulling that dialogue together. And when you think about the measures themselves you kind of look and say well, you know, $25 change on $5,000, you know, in a large sample size, well that's statistically significant. But is it really clinically meaningful? How does it shape or affect or alter practice?

And sort of when I looked through

the comments and I thought about the comments that I had made into the system those were things that were sort of coming to my mind as we went through it. So I could go through some of the comments that were made. I think everybody is sort of seeing what those look like. But it might be fruitful now just to open it up for a general dialogue.

DR. NELSON: Thank you. So we have several cards up. Why don't we just go up the row and then across. So Dolores. Or is it Jack? Jack.

DR. NEEDLEMAN: So, on the risk adjustment issue and the inclusion of dual eligibles what we've seen is it's a challenge because there are two components to that. One is we'd like to be able to look at SES because we're interested in discrimination against people of low SES, problems of access or higher use of some services because of delayed care and any number of other things. And if they're buried in the risk adjustment model

it's hard to see that.

On the other hand we have heard and seen documentation that the dual eligibles also seem to have more healthcare use than can be strictly attributed to their HCC status which would argue that there's something being captured clinically about that population it's important to risk-adjust for if we're trying to understand the differences in groups that have large numbers of, higher numbers of dual eligibles having higher costs. That you'd want to risk-adjust for that if those costs are due to their clinical situation. So that's the challenge we have.

And right now part of the problem with the risk adjustment model is that the ability to analyze that is buried in the model because the adjustment takes place at a very early stage. It is hidden in that percentage of expected -- the -- it's hidden in that percentage of expected cost figure that's subsequently used in the risk adjustment

model.

So at some point it would be good if one could simply get as an ancillary set of analyses what the dual eligibles look like on that measure and some other use measures so that one can analyze their experience separately to look for the kind of disparities that's the rationale for not putting the SES into the risk adjustment because we want to be able to analyze it explicitly.

DR. NELSON: Thank you. Brent?

DR. ASPLIN: Quickly on the exclusion of decedents is I think just a big policy gap. I don't know that just throwing them into the model beneficiaries who die during the year, that may complicate things. But either running a separate model or reporting both a version that includes decedents and one that does not would be important from a policy perspective given what we know about expenditures in the last year of life and how critical that is. So I feel

pretty strongly about that.

And then a quick question on the attribution. How closely does this mirror the attribution model in MSSP and Pioneer? Because you have 260 organizations in these now. I know it's close but it doesn't feel like it's exactly right. Or is it? That's my question.

DR. NELSON: Let's get an answer from the developers.

MR. BALLOU: I'm not sure if I'm understanding but in terms of mirroring, in terms of the construction logic it's the same. It's the same rule.

DR. ASPLIN: Identical?

MR. BALLOU: It's the identical rule. And that's a large part of why it was adopted.

DR. NELSON: Thank you. Nancy?

DR. GARRETT: So I think, Brent, I think you captured the issue before us very well here in terms of the situation we're in

with -- if we adjust for socioeconomic status in this measure then what are the implications of that from a broader perspective.

But I'm going to argue that I think we should. And so my thought process is that, first of all, based on coming from a safety net care system every day we see patients where their exposure to risk is different because of the vulnerability of their situations.

And so, you know, situations like a 23-year-old who is exposed to gun violence. A 40-year-old who comes in with advanced liver disease because of alcohol abuse. I mean these are real issues that they face because of their situation economically. And so it affects their exposure to risk but it also affects their ability to manage health.

And when we discharge a patient to a homeless shelter and they're kicked out from 8 o'clock in the morning until 6 o'clock at night and it's the middle of winter in

Minnesota that's dangerous for someone who got discharged with pneumonia. And so they're these social factors that are hard to measure but they absolutely affect health and our ability to take care of the patients.

And so while we need to be and we're working on lots of creative ways to address those situations for us to be compared on a level playing field with providers that have a very different kind of population I think that that is a problem with these kinds of measures. So I argue that I think we really should stratify it.

And in fact I think dual eligibility status is a good start but it's not enough. There's huge variation even within a Medicaid population in terms of access to resources and social factors like social support and all the other constellation of things that go along with social risk.

MS. WILBON: I'm sorry, I just have a question, Nancy. I thought I heard you

say you think that we should stratify by? Did you mean include it in the model or did you mean actually stratify it?

DR. GARRETT: I meant include it in the model. I'm sorry, yes.

MS. WILBON: Okay, I just wanted to clarify that?

DR. GARRETT: Yes. And one of the reasons I think that's important is because from a practical perspective the complexity of these cost and resource measures, it's going to be really hard to stratify afterwards in any way that's meaningful from a policy perspective. I think it has to be built into the measure in order for it to be really applied in a way that starts to level the playing field.

DR. NELSON: Cheryl?

DR. DAMBERG: I just want to make sure I'm understanding this dual status. Because I'm looking at the one table that shows the distribution of people with dual

status versus not across these groups by different group size and I don't see any variation.

DR. NELSON: What table are you looking at? In case others wish to view it.

DR. DAMBERG: Exhibit 1.2. It says, "Summary of characteristics of beneficiaries attributed to medical group practices for groups with at least 25 eligible professionals and at least 20 attributed beneficiaries by group size."

So the groups are 25 to 50, 51 to 100, 101 to 200 and more than 200. And I guess I was expecting to maybe see the duals congregating in certain group sizes but I'm not seeing that. So if they're sort of equally distributed.

Now, maybe they're not equally distributed at an individual physician level and that's something I can't tell from this table.

DR. NELSON: David?

DR. REDFEARN: I just want to throw one more punch at that poor horse and mention something that nobody else has mentioned so far. The risk scores that they calculate are composed of -- there's two different types of patients. There's the new enrollee, the community new enrollee and then there's the enrollees in the full one. The full one uses the HCC. The other one only uses demographic risk factors.

It's going to be a worse adjustment factor no matter how you look at it. And they're mixed in here. And you're going to have both of them going on and they're based on different structure. And that concerns me because there's going to be variation in how well that adjustment factor works depending on the status of the patient. That's an issue.

DR. NELSON: Would you like to respond?

MR. BALLOU: Yes, thank you.

That's certainly true and I'm not sure if Greg Pope is still able to join us on the phone. He may want to respond afterwards as well if he's still on the line.

The new enrollee model does not predict as well as you would expect, right. It doesn't have the comorbidities included.

Effectively what happens in our second stage risk adjustment then is we have new enrollees in a group are essentially compared to new enrollees in other groups. And continuing enrollees with the community data with the comorbidities in the same group are compared to continuing enrollees in other groups.

So I think that while it's correct that the precision will not be there for the new enrollees as it will be for those with the comorbidity data included we have not seen evidence to suggest that any meaningful groups is attributed a disproportionately large number of, say, new enrollees compared to

continuing enrollees as might happen when groups have, you know, they happen to get all the new Medicare beneficiaries that are just enrolling, for example. And I think that's where that would become more of a concern.

DR. NELSON: Daniel.

MR. WOLFSON: I want to take a contrary view about socioeconomic class. I don't think you should adjust for it. I think it gives providers a pass. And I can remember Patty Gabow from Denver Health who has a very difficult population who outperforms every hospital in the country on any quality measure. And she's set up services to do that.

If we let the socioeconomic class be a pass then why would somebody be motivated to have the proper services to deal with difficult populations? I would do something different and I would revert to percentage of improvement from one year to another. So I would adjust for where people begin in

resource use.

So I think there's other ways to do that. But to mask the socioeconomic, I think that's why people keep saying not to do that. And maybe -- well, we already have a two-class system. But we'll have a worse two-class system and it will just perpetuate the two-class system I think by adjusting for it.

And Patty Gabow's Denver Health is an excellent example of somebody who's overcome very difficult populations and outscores everybody in the nation.

DR. NELSON: It's 5 o'clock. About time to adjourn. Difficult area here, validity. We should I think take a vote. Before doing that, any public comments?

OPERATOR: At this time if you have a question or a comment please press *1 on your telephone keypad.

MS. WHEATLEY: Hi, Mary Wheatley from the AAMC again. We've been working with the academic practices looking at these

quality resource use reports and looking at the data and trying to understand the cost measure for the past couple of years.

I think there are a couple of things that I'd like to give from that perspective of trying to analyze and trying to understand this data.

One is this is -- we've actually not seen the data with this new attribution methodology. So it's new to the chief pro groups.

Before it was based on all specialty care and we actually thought that brought in a ton of noise, especially when you're pairing that with primary care measures. So we actually encourage going to a primary care attribution methodology but we haven't seen how that actually plays out in the real world in knowing how well this particular attribution methodology works. And if there needs to be additional tweaks.

And then the other thing that I

think needs to go with that is for a whole host of reasons we don't even know the full patient populations. Like the transparency that goes in with some of this measurement has not, you know, we know who gets assigned to the quality buckets that are kind of paired with these patients, but we actually don't know who the patients are.

And it's kind of this, you know, a lot of drilling down and trying to figure out what's going on with this data. And there's a lot of information that's not yet available. So I think that there's been a lot of progress made but there's a lot more that really needs to be done to really make sure we know that this attribution is right.

And then the last thing I would like to say is I know this new attribution methodology is used with the MSSP ACO models so those groups may have more experience with these measures. But this program and how this measure is being used in this cost metric is

a little different because you're not comparing yourself to your own historical history, you're comparing yourself across different organizations. And so that makes it a different way to consider this attribution method. Thank you.

DR. NELSON: Any other comments?

OPERATOR: We have no public comments at this time.

DR. NELSON: Very good, thank you. Any further comments from our committee or from our measure submitters? CMS, Mathematica? Okay. Good discussion. Complex issues.

So this is an all-in vote on validity, one to four. And I think we're ready to take the vote.

MR. WILLIAMSON: We will now vote on overall validity. You have 60 seconds beginning now. And we have 12 moderate and 12 low.

MR. AMIN: And Larry's out of the

room.

DR. NELSON: So a second time. Do you want to clear the system?

MR. WILLIAMSON: We'll now vote again on overall validity. You have 60 seconds beginning now. We have 13 moderate and 12 low.

DR. NELSON: Is that a correct count of everyone that should be voting?

MR. WILLIAMSON: Yes, that's 25.

DR. NELSON: Okay.

MS. WILBON: So.

DR. NELSON: We've seen similar numbers today.

MS. WILBON: Yes, we definitely have. You know, same scenario with the last measure. We will -- we'll continue tomorrow. You guys have done a great job today. Thanks for hanging in there. It's been exhausting, I know.

Well reconvene tomorrow. Again, I think tomorrow we're going to put our heads

together. Whether this actually means there's consensus obviously, probably not, but just based on operationally we will continue to evaluate the measure and see how things pan out in the end of the recommendation.

And we'll reconvene tomorrow actually at 8:30. So the agenda I think for tomorrow that we sent out said that we would start at 9:30. Because we kind of went so long today and had an extended conversation, which is great, we are going to start an hour early tomorrow. So we will be starting the meeting at 8:30. So if you can come by 8:15, breakfast will be out by 8 but at least if you can plan to be here by 8:15 so we can start promptly at 8:30 that would be great.

Again, if you have any juice, anything left and you want to come and hang out for drinks we will be there. So feel free to do that. We've emailed everyone the address and the location and time for that if you'd like to join.

Thank you, everyone, for a great discussion today. And to our co-chairs for leading us through.

(Whereupon, the foregoing matter went off the record at 5:05 p.m.)

E

H

M

S

C E R T I F I C A T E

This is to certify that the foregoing transcript

In the matter of: Cost and Resource Use Steering Committee

Before: NQF

Date: 05-08-13

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