NATIONAL QUALITY FORUM

+ + + + +

RESOURCE USE CARDIOVASCULAR/DIABETES TECHNICAL ADVISORY PANEL MEETING

+ + + + +

TUESDAY

MAY 10, 2011

+ + + + +

The Technical Advisory Panel met at the offices of National Quality Forum, Suite 600 North, 601 13th Street, N.W., Washington, D.C., at 8:30 a.m., Jeptha Curtis and James Rosenzweig, Co-Chairs, presiding.

PRESENT:

JEPTHA CURTIS, MD, FACC, Co-Chair, Yale University School of Medicine
JAMES ROSENZWEIG, MD, Co-Chair, Boston Medical Center and Boston University School of Medicine

MARY ANN CLARK, MHA, Neocure Group
CONSTANCE HWANG, MD, MPH, Resolution Health, Inc.
THOMAS MARWICK, MBBS, PhD, Cleveland Clinic
DAVID PALESTRANT, MD, Cedars-Sinai Medical Center*
BRENDA PARKER, PharmD, GlaxoSmithKline

KATHERINE REEDER, PhD, RN, University of Kansas School of Nursing
WILLIAM WEINTRAUB, MD, Christiana Care Health System

NQF STAFF:

TAROON AMIN, MPH

HEIDI BOSSLEY, MSN, MBA

HELEN BURSTIN, MD, MPH

SARAH FANTA

ANN HAMMERSMITH, JD

SALLY TURBYVILLE, MA, MS

ASHLIE WILBON, MPH, BSN

ALSO PRESENT:

CARLOS ALZOLA, MS, Data Insights

BEN HAMLIN, MPH, National Committee for Quality Assurance (NCQA)

TODD LEE, PharmD, PhD, American Board of Medical Specialties (ABMS)*

TOM LYNN, MD, Ingenix

KEVIN STROUPE, PhD, ABMS-REF*

KEVIN WEISS, MD, MPH, American Board of Medical Specialties (ABMS)*

* Participating via telephone.

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

(9:00 a.m.)

MS. TURBYVILLE: Good morning and welcome, everyone.

I want to give Helen Burstin as well as the two Co-Chairs, Jeff and James a chance to welcome you at this time.

DR. BURSTIN: Good morning, everybody. Helen Burstin, I'm the Senior V.P. for Performance Measures at NQF. Thank you for coming to what I think will be an incredibly interesting meeting. This is our first foray into resource use measures. We'd had a brief one phone call with the steering committee, but this is our first in-person meeting. So we really do view the resource use measures as being critical building blocks toward getting to measures that get us at value. And Jeff has already had a measurement framework a couple of years ago that made it very clear that we don't believe these measures in isolation -- should be used in

isolation, and should always be coupled with quality measures to get at value. So we really do view these as being important building blocks that will ultimately help us knit those together in better measures of efficiency and value.

So we've got a great team working with you today, and thank you.

CO-CHAIR CURTIS: Hi, Jeptha Curtis. Thanks to everyone for all the time you've already put on this project in advance of all the time we're going to do in the next two days. I think it's going to be a very intense kind of meeting with something of a process, potentially an iterative process where, as we work through this new process or this -- these new sets of measures.

This is the first time, as Helen said, that resource use measures have been evaluated by the NQF, and we are the first TAP within the resource measure process. So this is the first time that these criteria have

been implemented, which is challenging. And so we are -- maybe it's not the right analogy, but I think we're all, to a certain extent, going to be feeling our way in the dark and relying on each other to bring the issues forth.

The advantages, that there are, I think three measures, measure developers that are represented in this set of measures, and so within each, there's seven ABMS measures that I think all have very common themes to them. So as we're discussing the first one, I expect that we'll probably end up spending a lot of time going through that. And once we get that traction, that will help us facilitate us going through the rest of the measures.

I think the one thing that I want to be aware of or make you guys, something to be cognizant of is, there's this possibility of drift in our evaluation as we're going through them, that the criteria or the

threshold that we're setting for the first measure as we're going through it may change over time. And so I don't think there's a formal way of being aware of it, aside from having -- asking the NQF staff to sort of keep us on target and potentially bring us back to measures if they get the sense that we've drifted away and are applying a different threshold than we did originally.

And in terms of keeping us on timelines, I think we will be using the parking lot with some frequency for some issues that we don't think we'll be able to resolve in a timely fashion, but that are important enough that they would need to be fed back to the measure developers or addressed in a group format going on.

So with that, I'll turn it over to Jamie.

CO-CHAIR ROSENZWEIG: Yes, hi, I'm Jamie Rosenzweig. I'm co-chairing this meeting along with Jeptha, and I'm -- whereas

Jeptha is a cardiologist and most of these measures relate to cardiology, I am an endocrinologist and my focus has always been related to diabetes, although to a large extent I've done a lot of work related to cardiovascular risk in diabetes.

And so I will be focusing at least more on the diabetes-related measures and also the comorbidity aspects for the -- that apply to the processes that we'll be reviewing.

I have a background in quality improvement measure development, cost and resource uses to a certain extent is a newer field for me, although I've done work related to costs -- disease management programs and the costs and how they affect the costs of care. So this is going to be a very interesting process. And although I think we want to be able to keep on schedule, I think since we're the first of various groups that are going to be dealing with these particular types of measures for NQF, we need to make

sure that we cover these areas adequately. And I think what we'll -- you know, we'll have to do the best we can to be able to fit -- to be able to take care of as much as possible during the time allotted.

I would mention that we have excellent vendors here who are providing very interesting ways of looking at these processes, and the ways of looking at them are very interesting and also very different. But they're not only -- they won't only -- one particular process may apply more effectively for certain disease states than they will for others. And some of the disease states we're talking about have many more acute related problems, and others are much more chronic and less self-limited in terms of times of episodes of care.

So we're going to have to really look at each disease state independently to a certain extent and see how they apply to these particular approaches.

And I want to thank everyone for their extensive efforts on really a -- going through a tremendous amount of material in order to prepare for this meeting.

MS. WILBON: And while everyone's kind of going through introductions, we have our general counsel, Ann Hammersmith, here. She's going to walk us through the disclosure of interest process. So since this is our first time evaluating measures, we do need everyone to kind of go around and disclose any conflicts of interest and maybe as you call their names they could -- just allow them to give a brief introduction of themselves and disclose their interests at that time, as well.

MS. HAMMERSMITH: Good morning, everyone. As you recall, we sent you a disclosure of interest form and policy several weeks ago, maybe a month ago even, which you filled out. And we vet these very carefully, so if you're sitting on the committee, it's

highly unlikely that we regard you as having a real or apparent conflict of interest.

But in the spirit of openness, we're going to ask you to go around the table and disclose any interests that you believe are relevant to your service on this panel, in particular any grants or consulting arrangements that you have that are -- that you believe are related to the material before the committee.

What I'd ask you to do is go around the table, identify yourself, tell us who you are with. You don't need to recount your resume but obviously we are interested in what your background is. I also want to remind you that you sit on this committee as individuals. You do not represent the interests of any group, including your employer or any group that may have nominated you to sit on this committee.

So with that, I will turn to my right and start with you, Ms. Reeder. NQF

staff, you don't need to disclose anything.

DR. REEDER: Hello, my name is Kay Reeder, from Kansas University Medical Center in Kansas City, Kansas. I have an NIH five-year K99 R00 awarded August 2010 through July 15, and I'm a consultant for a cardiology fellowship in Iowa, and -- remotely. And I do work with the Iowa Health System based in Des Moines. I have no conflicts of interest.

MS. PARKER: Hi, good morning, Brenda Marie Parker. I often say Marie, so that's why I include that. I have a background in pharmacy and a Master's in Public Health and Health Outcomes. I've managed patients with diabetes and cardiovascular conditions, mainly chronic conditions. I work for GlaxoSmithKline in the applied outcomes group there and have no conflict of interest.

MS. CLARK: Hi, I'm Mary Ann Clark, I work for a small health care consulting firm, Neocure Group, based here in

D.C. We work with medical device manufacturers on health economics and reimbursement-related issues. I have a lot of experience in the cardiovascular area. I formerly worked for a Boston scientific corporation evaluating technologies there. And many, many years ago, I worked on the Harvard RBRVS Study with Bill Hsaio. And in terms of conflicts, because I work for a consulting firm, we have a lot of cardiovascular device companies as clients, large and small. So that's my disclosure, I guess.

DR. HWONG: Hi, good morning. I'm Connie Hwong, I'm the director of Clinical Affairs and Analytics at Resolution Health which is a wholly owned subsidiary of Wellpoint. I'm a general and internal medicine physician. My experience with quality measures is we do, at Resolution Health, a lot of physician quality profiling. We also are measure developers and we have 25

NQF-endorsed clinically enriched claims-based quality measures which are primarily process measures.

The physician quality profiling we do is also coupled with efficiency scores, so that is some of the sort of working experience I've had in terms of the applications of quality measures and efficiency. I have no conflicts to disclose. Thank you.

CO-CHAIR ROSENZWEIG: I'm James Rosenzweig. I'm a director of Diabetes Services at Boston University School of Medicine and Boston Medical Center. I am the -- with respect to conflicts of interest, I am the Chair of the Performance Measures Subcommittee of the Endocrine Society and also member of their Clinical Affairs Core Committee. And I represent the Endocrine Society at the American Medical Association Physician Consortium for Performance Improvement.

I was the former Chair of the

National Diabetes Quality Improvement Alliance, which was the original group that put together a set of quality and performance measures for diabetes that were eventually submitted to NQF. I've been on a number of -- I've been on two other, I think, diabetes Technical Expert Panels in the past. The one last year related to episodes of care.

With respect to specific conflicts of interest, I'm on the Scientific Advisory Board of Alere Medical, which is a disease management company, for which I receive a small honorarium. And I have a -- I've been involved in several educational programs through Boston University that are supported by unrestricted educational grants for CME-related activities from several organizations, including the Hearst Foundation and Sanofi-Aventis.

CO-CHAIR CURTIS: Jeptha Curtis, I work at Yale University in the Yale Center for Outcomes Research and Evaluation. I am a

clinical cardiologist and interventional cardiologist as well as a health services researcher.

My experience in quality metrics in general is -- has been in the development of outcomes measures. I've been part of the team at Yale that has developed the six publicly reported outcomes measures for AMI, pneumonia and heart failure, for better or worse. And I have more recently been involved with additional measures for PCI mortality and readmission, none of which I think represent a conflict of interest for this particular endeavor. And -- with the exception, I guess I do receive salary support under the CMS contract for measure development.

That should be it.

DR. WEINTRAUB: Good morning, everybody. I'm Bill Weintraub. I'm Chairman of cardiology at Christiana Care in Delaware, professor of medicine at Thomas Jefferson University and professor of Health Sciences at

the University of Delaware. I was at Emory University for many years, and I'm professor emeritus of medicine and public health as an investigator. I'm a cardiovascular epidemiologist. I've had federal funding for the last 30 years and hope to continue doing that for the next 30 years. We'll see.

I have been very involved with the American College of Cardiology and the American Heart Association. I was one of the people that developed the National Cardiovascular Data Registry and remain on the Registry Board. I'm also on the informatics committee and I'm the incoming Chair of the Data Standards Committee for -- or a task force, really, for -- which is a combined task force of the American Heart Association and the American College of Cardiology.

At the AHA, I'm on the advocacy committee and I'm the incoming president of the Great Rivers affiliate. I have many relationships with industry, all of which I've

disclosed, but I don't believe any of them are conflict of interest for these activities. Thank you.

DR. MARWICK: My name is Tom Marwick, I'm a cardiologist at Cleveland Clinic. My interest is in cardiovascular imaging and particularly in outcomes research related to that. I have a number of grants related to technical developments with industry, but none that are pertinent to this activity.

MS. HAMMERSMITH: Thank you for those disclosures. Are there any panel members on the phone?

DR. PALESTRANT: Yes, my name is David Palestrant. I'm a stroke neurologist and neurointensivist at Cedars Sinai Medical Center. I built both programs here at Cedars so I have some background in performance metrics.

I have -- my disclosures are I'm OPI and PI on numerous multi-center studies

for which I receive no direct funding. And I have no other conflicts.

MS. HAMMERSMITH: Okay, thanks.

Is there anyone else on the phone?

(No response.)

MS. HAMMERSMITH: Do any of you have any questions of me or anything that you want to discuss with each other based on the disclosures that have been made today?

(No response.)

MS. HAMMERSMITH: Okay, that's the usual response. Thank you. Have a good meeting.

MS. WILBON: Thank you, Ann. So now that everyone on the TAP hopefully is a little more familiar with each other, staff, I guess we should introduce each other. I think we had opportunity to hopefully greet each one of you as you came in but we'll -- I'll start with Sally and we can introduce ourselves.

MS. TURBYVILLE: Good morning,

everyone. I am Sally Turbyville, and we have all met via phone. I'm very pleased to see everyone here today. I'm the senior director on this project working along with the rest of the team who will introduce themselves. And we are really looking forward to today. We acknowledge that there are a lot of materials involved and that this is new, so we expect some bumps along the way. And we look to you to understand if there's anything we can do to improve the process in real time or looking forward as we continue to work with all of you through this project. So welcome.

MR. AMIN: My name is Taroon Amin. I am assisting the team as a senior director on this project, as well. I will be tasked with the effort of making sure that the measure evaluation drift is minimal. So I look forward to the discussions over the next two days.

MS. WILBON: So good morning. I'm Ashlie Wilbon, I'm the project manager on the

project. And it's nice to finally put faces to names and, yes, I'm the person who's been sending all those emails to you. So thank you for your patience. And it's been -- it's a new process for us all and we're trying to do our best not to inundate everyone and realizing that it's a lot of information. So thank you, everyone, for reviewing everything and for being here today.

And I'll be just taking notes through the process and making sure that we've captured everything throughout the course of the meeting and making sure we're sticking to our process. Thanks.

MS. FANTA: Good morning, everyone. I'm Sarah Fanta, Research Analyst, NQF, working on this project with the rest of the team. I'm really looking forward to working with all of you during this process.

MS. WILBON: So I think we'll start with the folders that everyone got. I just want to walk you through what's in that

folder. There will be several materials in there that you might want to refer to throughout the day.

So the first paper in the right side of your folder should be an agenda, followed by, I believe, either -- I took -- I moved some of my papers around, so it's either the slides that we're going to go over or the roster, followed by the measure review assignments that we sent out. So each measure is on the left followed by -- next to the assigned reviewers and the lead discussant is in bold. So that will be -- we'll talk a little bit more about what's involved for the lead discussant in a few minutes. But that list for you to refer to.

That is followed by a table of the submitted measures with the title, description and developer. And that should be followed by the actual resource use evaluation criteria, which includes the notes in that packet. That is followed by the table of -- the side-by-

side table of the criteria and the measure submission items that we sent out. I suspect that, as we start getting through -- getting into the actual subcriteria ratings, that that will be very helpful to pull out.

That should be followed by the summary that we sent out on Friday. We've compiled all of the online measure evaluations that had been submitted as of, like, Friday evening and we sent that out. And then we did the same thing last night. So the document following that is what was submitted as of 5-9.

So we were hoping that, for the lead discussants, as you're introducing the measure and summarizing what's been evaluated so far, that you'd be able to use that information to kind of give everyone an idea of, you know, where people agreed or disagreed. And what people's general feeling was about the measure for those who submitted their evaluation.

And then followed that -- following those two documents are the travel memo that we sent via email along with the reimbursement form. It's on paper. We sent you the Excel file as well, so as you're here, please remember to keep all your receipts. Meals should be itemized receipts and you can either write it down on paper or, you know, type it in to the Excel spreadsheet when you get back home and send that in when you do your -- at the end of the trip when you submit your reimbursements to NQF.

So you should also have -- I'm sorry if I missed it -- is a table that has a side-by-side of the reliability and validity -- evaluating reliability and validity. So we'll be referring to that, as well. It's a cheat sheet to kind of, as we're evaluating scientific acceptability as to the types of things you should be looking for, as you're rating things high, moderate or low. So again, it's another piece of information to

add to your already exploding brains, I'm sure.

All right. So I'm going to -- we're going to go ahead and just do a quick introduction to the meeting and hopefully get out of the way some of the process-oriented things and goals for the meeting and what we'll be looking for today.

So our agenda, we've already done our roll call and DOI and we're going to get into some of the goals and objectives for the meeting.

So we've mentioned already a couple times today that, you know, one of our main goals, in addition to understanding the consensus-development process, we'll do a quick review of that, really understanding the subcriteria-evaluation process. And as Jeptha mentioned, it's going to be sort of an iterative process. This is the first time we've applied the criteria that we've developed for resource use measures to the

resource use measures in this process.

So and as we're beginning the measure review process, we'll start with the measure overview, which will be started with the measure developers giving a brief overview of the measures and then the lead discussant jumping off the discussion for the TAP.

That will lead into the discussion of the measures by each of the subcriteria, and then we'll have you rate each of the subcriteria with using your remotes. And those will be the final ratings that we will use and forward on to the steering committee who will look at what you guys have discussed and how you've rated the subcriteria. And then make their recommendations on the overall criteria and ultimately the measure.

And again, we're -- you guys are the first group. You're our guinea pigs, so to speak, so throughout the process we're going to be looking to find ways to either find efficiencies or improve as we go along.

So with the consensus-development process, we start with the project-specific topic. This project was funded by HHS and it's been going on for a while. But for this consensus-development process, we're actually reviewing the measures. Once we have the project topic, we gather the steering committee and the TAPs to review the measures and give us their expert opinions on -- based on the criteria.

Once we've got those recommendations, we pull together a draft report and we put that out for public and member comment. And after the public and member comments on that, we bring it back to the steering committee, usually to provide any inputs. Sometimes it may change some of their, you know, views on how they have evaluated the measures.

We then put it out for member voting. The public and member comments, the member votes then go to the CSAC for review.

And our CSAC, or Consensus Standards Approval Committee, is an overarching body that reviews the recommendations of all the committees and TAPs that -- for each of our projects to make sure the process was adhered to and that the recommendations that were made should move forward to the Board for ratification.

And once the Board has ratified the measures, we put those measures out for appeals in case anyone has any final comments about the measures. So that's, in a nutshell, a high-level overview of the consensus-development process. And again, we're focused right there where the yellow and the red circles are.

So for this particular project, we've divided into two cycles. And these next few slides you've already seen before, but we figure repetition is the best way to approach it at this point; there's so much information. So for this first cycles, we're only looking at the cardiovascular diabetes and non-

condition-specific measures. So we tried to take off a smaller chunk to start with, realizing this was a new process, and learn from the process for this first cycle, and hopefully apply any things that we've learned towards the next cycle, which has pulmonary, cancer and bone joint measures.

So the measure review process, we start out with the staff review, once the measures were submitted. So even though you guys got the measures about two and a half weeks ago, we've been working with the developers the last two months to try to get the submission forms to a point where they were complete enough and responsive to the question and to a point that we felt like was good enough to pass on to you guys to begin review. So there's been a lot of work before you get that to make sure that it's ready to pass on for review.

We just obtained -- yay - our statistical consultant who will be attending

the meeting today, starting at about 10:00 a.m. to 2:00 p.m. today. He's really going to be focusing on the risk-adjustment methodology and the testing -- reliability and validity testing information that was submitted with the measures. We've asked him to focus on each of the measure developers, based on kind of what Jeptha was saying, that a lot of the measures submitted by an individual developer, they used kind of the same methods throughout each of the measures, regardless of what the focus of it was. So we're going to have Carlos come and answer any -- share with you what his analysis was of each of those developers' methodology, so you have an opportunity to ask questions. And hopefully what he's able to share with you on those first three developers will help you carry through for the rest of the meeting those principles to help you apply going forward.

So the -- on the -- I guess the left side, the steering committee has already

started reviewing the non-condition-specific measures. We had a conference call with them last week, I believe. And then you guys are on the right side with the TAP evaluation of the condition-specific measures, starting out with cardiovascular. And as I mentioned, what you guys -- however you guys -- the ratings that you guys submit today will be passed on to the steering committee for their evaluation of the overall criteria and recommendations for endorsement.

So the role of the TAP at this point is to evaluate the candidate measures against the NQF evaluation subcriteria to identify the strengths and weaknesses of the measure focusing on the clinical logic and provide guidance, again, to the clinical applications of the measures. Again, the composition of the TAPs is quite different from that of the steering committee. We purposely kind of made the TAPs really heavy in methodologists and clinicians who are

focused on the condition area of the measures.

So what we're going to be trying to focus on today as we're evaluating the measures, we understand having gone through our first call with the steering committee that there's going to be a period where everyone's just trying to understand in general the concept of the measures. So you know, kind of ask any questions of the developer that needs to go on so everyone's comfortable with taking a deeper dive into the measure. And once we're ready to do that deeper dive, we're really going to try to keep it focused on the criteria.

So ultimately what we need you guys to do is feel comfortable with rating the criteria based on what was submitted for the measure on a scale of high, medium or low, so that that can be passed on to the steering committee.

And we'll -- also in your folder, you should also have a paper with the voting

instructions on there. And I think before we do our first voting using those remotes, we'll kind of go over the mechanics of it. But essentially, Sarah will be controlling the slides over there and we'll just ask you to hit 1, 2, 3, 4 or 5, based on the subcriteria that will be on the screen. You just hit your number, and it -- the results will show up on this little -- this big screen over here to the right. And that's what we'll be capturing for the steering committee.

Again, this is just a high-level timeline for Cycle One. We're obviously at the CV/Diabetes TAP meeting today. We have a steering committee meeting coming up at the end of June. We've actually just scheduled a follow-up steering committee call on June 6th to continue their discussion of the non-condition-specific measures.

The goal of Cycle One is to have some -- some measures endorsed by the end of this year, so realizing that this is our first

time, again, we wanted to chunk it out so we end up with something by the end of the year.

Cycle Two starts with our first TAP meeting at the end of June as well, and we're hoping to have those -- some of those measures through the endorsement process by March of 2012.

And I'm going to go ahead and pause there and ask for any questions, and then hand it over to Sally to go into a little bit more detail about the evaluation process. Does anyone have any questions or --

DR. MARWICK: How will you use the scores?

MS. WILBON: The scores that you guys are submitting today: the high, medium, low? So what we do is we compile that and do a document along with the rationale that you guys associate with each of the ratings. So if you rated something high, we're going to be looking for you to explain why you think it should be rated high. That information goes

on to the steering committee, and they use that to determine overall whether or not the measure should be recommended, so they're not starting from scratch.

A lot of the steering committee members aren't clinicians -- there are some clinicians on there, but there aren't a whole lot of clinicians. So they're really going to be looking to you guys for that guidance on those parts of the measure to make a broader recommendation on the measure to move forward through the process.

Any other questions before we move forward?

(No response.)

MS. WILBON: Okay, thank you.

MS. TURBYVILLE: Thanks, Ashlie.

So we want to spend a little bit of time talking to you about the evaluation process. But first, before we get started, we want to make sure that we're all on the same page in how the steering committee defined

resource use for this project.

As Helen mentioned, there was an early-on acknowledgment that resource use measures by themselves are not measures of efficiency, but that they're really important building blocks to contribute to that understanding. And so NQF made a decision to embark upon an endorsement process for resource use measures.

And in that effort, last year when we were working with the steering committee, we wanted to make sure we had a common definition in moving forward. And basically there are measures that compare health services counts, and they can be in terms of units, so frequencies, or they can be monetized so it can be a standard dollar applied or an allowable charges, et cetera. Really it's up to the measure developer to determine how they want to count the services and then they should provide the context and the reason for that approach.

So we are going to ask you today, as Ashlie mentioned, to evaluate and rate the submitted measures as they're specified against the NQF resource use evaluation subcriteria. And as we go through this process, it will be by measure, and we will be looking for you to rate against the evaluation subcriteria sequentially. And so as we get into it, I think the process will be a little bit clearer.

So as you know, there are four major subcriteria, and they are importance to measure and report, scientific acceptability of the measure properties, usability of the measures and whether or not it's actually feasible to implement the measures as they're specified.

We also, while we did not ask the measure developers, because this is the first resource use project, to attempt to harmonize their measures against any existing endorsed measures. As we go through the process, there

may be, for example, age bands or other areas where we may ask measure developers to go back and harmonize. And that's basically looking for them to have some amount of similarity. It may not come up, and if there's a decision to not attempt to harmonize, we'll be turning to the TAPs or the steering committee to justify those particular decisions.

So I want to go into a little bit detail about the subcriteria, since that's what all of you will be focusing on for the measures. And I have also borrowed extensively from the testing task force report that we sent to all of you earlier on. But it's quite lengthy, and given that there's a lot of materials of the measures to review, I wanted to use this as an opportunity to summarize it the best that I can so that you have a nice, robust set of tools as you move forward.

So the first area is about the measure's focus is important. So the first

criteria doesn't necessarily get into how the measure is specified being important; it's have the measure developers chosen an area, topical area, that it's important to measure area resource use. And some of the ways the subcriteria support these decisions is if they're looking in an area that's a national health goal or priority area or high impact. And it's also, is it a problem area.

We also ask you to evaluate and rate whether or not the purpose and objective that they have submitted is clear and resonates well with the topical area that they have chosen. And certainly whether or not 1-D subcriteria is about whether or not, given the area that they're measured, the resource use categories that they've selected: does it make sense?

So we will be asking you to rate those subcriteria as far as it being an important area to measure.

Scientific acceptability is where

we first start to dive into as the measure is specified, in trying to think through the reliability. So is the -- are the results consistent and potentially consistent when nationally implemented, and validity. So is the measure, as specified, credible?

They have these two, the first two, 2.A and 2.B actually have a lot of sub-subcriteria, if you have. So in order to assess reliability, there are actually two subcriteria under that. And then validity is the one that's quite lengthy with six sub-subcriteria. If you can come up with a better word for that, I'm open to it.

But anyway -- so in thinking about what the task force recommended in their report, they did recommend that empirical evidence of reliability and validity should be expected for all measures endorsed by NQF. And certainly for resource use measures, given the high complexity already, the steering committee and NQF agreed that, in order for a

measure to be evaluated, there must be empirical reliable and validity testing. But we'll get into validity and talk about the exceptions that are made there.

So although the testing task force also recognized that, although reliability and validity are not static properties and can really vary under different conditions of implementation, for example local practices of coding, structure of the data platforms, the purpose of the testing for NQF endorsement is to demonstrate that a measure can be reliable and valid when implemented as specified.

So we have to have a jump of faith that people will implement them as specified. We know certainly in the real world there is a tendency maybe to tweak that, but it is -- our charge is to think about the measure as specified.

So while implementing and reporting the measure is expected to lead to improvements in documentation, data coding and

data capture and thus improvements in reliability and validity -- this is an important point -- the assumption of approved reliability and validity over time applies to all measures.

We know this as a fact; once we start reporting, we expect those improvements to happen. It doesn't negate the need to demonstrate reliability and validity during the time of our endorsement consideration. So we can expect that the reliability and validity testing will have some limitations, and certainly we can expect reliability and validity to increase once it's implemented. We still will rely on you to determine how the reliability and validity testing is presented today, where the measure is now and if it meets the subcriteria.

So as I said, there are two sub-subcriteria under reliability. The first is whether or not the measure is clearly and precisely specified in a way that it will be

implemented consistently once endorsed, or if endorsed.

The other subcriteria under that is about repeatability of the measure data or the measure score is precise. And so there's two ways that they can meet 2.A.2.

So evidence of reliability can be accumulated over time, so NQF does allow the measure developers flexibility in how they want to demonstrate the reliability of the measures. And the scope of the testing may be relatively small in scale for initial endorsement. We do expect further analysis once a measure that is endorsed comes up for maintenance review. The reliability and validity testing would be expected to be at a higher bar. So this first initial endorsement process, the testing task force acknowledges that the scope of reliability and validity may be limited.

It's also important to note that reliability and validity testing may be

conducted for either the data elements -- so the data elements on which the measures rely on to run -- or the measure's score as it is computed. And this will have implications as you walk through considering and weighing in whether or not they're meeting that. And we'll get into how that's rated in a minute. In fact, that cheat sheet that has the reliability and validity goes -- has a cross-walk in the difference between data element reliability and validity and measure score, one being preferred potentially over the other.

So that's it for reliability.

Precise specifications, the measure is demonstrated to be repeatable. So for the validity, there is a lot more sub-subcriteria to think about. First, starting with whether the measure specifications are consistent with the evidence that they presented in the important section, particularly under criteria and subcriteria in

1.B. So, you know, they provide us information of what the purpose and the goal of the measure is under importance. At this point, as the measure is specified, is this consistent with what they said the purpose of the measure are.

So for example, if they were saying that it's important to measure the cost of care for diabetes, and then the measure as specified is about knee replacement, you may want to think about whether or not that meets their purpose. An absurd potential example, though just trying to point to what we're looking for.

The other 2.B.2 is also, again, getting to the data elements are correct or the score reflects the costs of care. It also asks you to think about whether or not the measure score can distinguish from higher and lower resource use.

Validity testing of data elements typically is about agreements with another

authoritative source of the same information. It can be from both published and unpublished sources. So what kinds of testing or validity assessment of the administrative databases are occurring. And it can also include systematic testing of face validity. And so that is an adequate validity assessment. It comes in as a lower bar, but acknowledging that some of the empirical analysis for validity testing, especially in this initial endorsement, may be beyond what is able to be done. If the measure hasn't been implemented much outside of a database, it may rely heavily on that face validity assessment.

There is also other types of validity testing. We are, as Ashlie said, having a consultant to help us support all of you in determining if the testing that they've used is adequate. But things like looking at the computed score against another measure that is considered valid, or looking at the correlation's relationships of that measure

score with something, another measure that is looking at the same thing. So there are different approaches in assessing validity important to think about.

Again, we will -- unfortunately the consultant just started on Friday, so I mean, he's done a very good job in assessing six of the measures, but -- and will be here for question and answers. But moving forward, it will be a lot -- you'll get that input much earlier on. So we would have liked to have gotten it to you a couple of weeks ago.

So data analysis -- moving on to the other two criteria under this -- is looking at demonstrating that the methods for scoring of analysis allows for identification of statistically significant results, and probably in this situation, practically meaningful differences in performance. So you know, even if it's -- seems like a small difference, is that meaningful to those who are measuring?

And then also, the potential for evidence of overall less than optimal performance. So is there evidence that, somewhere out there, there is less than optimal performance?

This last element, 2.B.6, I want to talk a little bit about, so it talks about the comparer results are demonstrated when there are different data sources being used. And primarily this gets at to when there are different options of data sources. So certainly we have acknowledged the difficulty of stitching together different data sources. But what this criterion and subcriterion is really getting at is, if there is an option to use different data sources in replacement of each other to produce the measure.

So for example, a measure developer may provide someone the option of computing a clinical target area, let's say a diabetes population. Using administrative data, or if they think their administrative

data isn't very complete, they may say you can go to the medical record.

We would expect measure developers to have tested those different options to see that they're coming up with comparable results. This is a little bit different of what we expect to see in resource use measures where we expect and probably want them to be using pharmacy data to calculate -- to include in the resource use estimation as well as the inpatient claims data, as well as the ambulatory claims data. We wouldn't expect the ambulatory claims resource use to be comparable to the inpatient resource use because they don't actually represent the same costs.

So I just want to make sure you understand the distinction of what this subcriterion is really getting at.

Any questions?

CO-CHAIR CURTIS: Practically, are there any examples where this is relevant in

this --

MS. TURBYVILLE: I didn't see any. I did not -- I mean, I've looked at most -- at least a couple from each vendor. I did not see any of them providing an option to go to the clinical record or EHR option or anything. It's -- the measures I've seen are all administrative-based and include different sources of that administrative data. But that's really to pull in the different costs and it's not options. It isn't you can use this or that.

DR. MARWICK: I have a question as well. So is your expectation that this information will be obtainable at some stage in the future or is it obtained now? It's just that, in terms of specifics, I don't see much evidence of this material in the documentation that I've looked at. I've seen generic statements about how it might be obtained, but not actually a data set that I could compare.

MS. TURBYVILLE: Right. So most -- I would imagine most of the developers didn't present any information to support this because they're not providing options of which databases to use in replacement of each other. But I think this -- I might be understanding correctly, this gets into usability and who the audience and the users of these measures are. Am I getting that right? So how would they obtain the data necessary to run the measures?

DR. MARWICK: Well, I think it's a bit before --

MS. TURBYVILLE: Okay.

DR. MARWICK: -- usability, really. I think it relates to some of the material that you're talking about already.

MS. TURBYVILLE: Okay.

DR. MARWICK: I mean, what I struggle with -- and I'd be interested to hear what other people on panel say -- is that although they talk about generically how they

could be used, there's no example of it actually being used to -- you know, for example, to understand the impact of different levels of risk and so on.

So what I'm having difficulty understanding is, are we voting today on whether this is something that's feasible or are we voting today on whether they've actually achieved this target?

MS. TURBYVILLE: So very good question. Depending on -- some of the measures are currently in use. Others, like AMBS-REF, they've developed them, they've tested them in databases, but they haven't -- and they acknowledge this in their submissions -- they haven't been implemented in a broad manner.

The testing that they did on database is acceptable for us to review at an initial endorsement -- as I said, the testing task force acknowledges that the scope of the testing may be limited in the -- may be

limited in the initial endorsement process. We would expect any measures that get endorsed today, when they come back for maintenance, which at minimum would be within three years, that they would provide more data to support how it's being implemented.

Now, to the extent that there is an assumption that users will have access to the data necessary, yes, I think in order to follow the specifications as specified, someone who wanted to use one of these measures would need to have access to the data that would support the specifications. If I'm answering your question correctly. And please, anyone on the TAP, if you -- yes?

DR. WEINTRAUB: Let me explain a little bit more on Tom's question, because I'm troubled by this, too. You know, the measures are real -- these look like a good idea, but the developers themselves make it clear that this is -- they're not really ready for primetime.

So I guess the disconnect is, if it's clear that they've developed something but the testing of it is really fairly minimal, I don't see how we can be asked to endorse it. I don't quite understand what we're being asked to do if the developers themselves feel it's not quite ready.

MS. TURBYVILLE: I'm not sure that the developers -- I -- don't think it's ready. I would only hope that they're submitting the measures for endorsement if they do. I think some of them acknowledge that they're not at the point where they have had time and opportunity to implement them broad. But the NCQA measures, for example, and the Ingenix measures are currently in use widely. So --

DR. WEINTRAUB: That's not the ones that I mean.

MS. TURBYVILLE: Okay. Well I think as we get into them measure by measure, if you see some of that, that is definitely important to bring to the attention to

everybody. But that's -- now, to the extent that you have concerns about the reliability and validity findings or the testing approach, absolutely we expect you to rate those and have your -- bring your expertise to this table. It's just important to acknowledge the guidance of the tasking force. It's not to take away from the work of this group.

CO-CHAIR CURTIS: So just to --

MS. TURBYVILLE: Please, Helen.

CO-CHAIR CURTIS: I'm sorry, Helen.

The -- we're not accepting promissory notes in this case, right? We're evaluating on the evidence we have in front of us. And to the extent possible, if they don't meet up -- and I think usability is kind of where a lot of these are falling down. We just accept what they have and move on. But again, we're not extrapolating or imputing.

DR. BURSTIN: And I'll just add that sometimes things may not feel ready for

primetime because they aren't actually in use. But we do accept testing of reliability and validity, even if it's not in widespread use as at least evidence for reliability and validity.

I think I have heard from -- there's some confusion. Some of them, for example, indicated the measure is probably for quality improvement but not potentially for other accountability functions. Those are the important questions to query the developers, and that's why they're here.

MS. TURBYVILLE: Okay. Another one of the subcriterion under validity is about disparities and whether or not any identified disparities are then addressed in some kind of stratification approach for measure scoring. We can talk about how relevant that is or is not for these measures, and we look to you to provide that guidance.

So usability, I think, certainly is going to be, as it always is, an important

criterion. And it's really thinking about who the intended audiences are and the intended users. And can the measures and the results, in particular, of the measures -- this really focuses on the results as well -- would they support decision making.

So one of the things we're looking for, are results reported to the public? Certainly we acknowledge, and there are exceptions allowed, that they're available to the public. So maybe they're not posted on the public web, but there is an approach to make sure people are receiving the benefit of the results. Are the results meaningful? Are -- can the audience, the intended audience, understand them, and will they be useful for public reporting and quality improvement?

We're looking for transparency and understanding of the supported measures. So for example, for someone who is being measured by these measures, are they able to understand how they're being measured?

And then we also spoke a little bit about the harmonization. We're not sure how that will pan out in resource use, and we'll continue to support the experts as we review the measures to look for or make -- provide rationale for that particular sub-sub-subcriteria.

And then feasibility. Are the data elements in which are required, or the data sets, to run the measures routinely generated? Are they generally available? Are they available electronically? I think all these measures are built on administrative data, so I suppose the exception would be if they're looking for some administrative data source that is rarely available.

And then also thinking about susceptibility to errors of the measures. Any unintended consequences of the measures themselves and kind of weighing in if those errors are unintended consequences or inconsequential themselves. Or at least can

be minimized or monitored. So we may acknowledge that there are some unintended consequences but, you know, if they can be monitored or minimized, is that something that the steering committee can rate on.

And then also certainly that the measure is implementable as specified.

Yes, please.

CO-CHAIR ROSENZWEIG: Going back to the previous slide, 3.B -- oh, I'm sorry. Going back to the previous slide, 3.B, when you're talking about harmonization, are you talking about harmonization of the particular measure sets here with the other measure sets here, or are you talking about harmonization with the quality measures that NQF has already endorsed?

MS. TURBYVILLE: Good question, thank you.

We are not talking about harmonization with the quality measures. Because we have not done a resource use

project, the only harmonization possible for this effort, and based on where we are now in thinking about efficiency in general, would be harmonization against each other, the measures that have been submitted under this project, if there's any that is applicable.

So we are not looking for the measure developers to harmonize against endorsed quality measures.

CO-CHAIR ROSENZWEIG: Thank you.

DR. BURSTIN: But at the same time, though, I think -- as I'm thinking about a diabetes example, for example, Jamie, if it would seem very strange, for example, to combine insulin-dependent, non-insulin-dependent diabetes on a quality measure. I think those are sort of lessons you might want to bring to this, even if they're not directly measures to be harmonized.

MS. TURBYVILLE: Right.

DR. BURSTIN: As an example.

MS. TURBYVILLE: So any questions

about -- I know I just threw a lot at you. But my hope is it will help throughout the next couple of days. At least you'll have them also in the slides to work through.

I wanted to spend a little bit of time talking about the measure modules and how we came up with them. Really, we came up with five measure modules to allow us to accommodate for the different types of resource use measures that we expected to see. And it was really for the purpose of collecting the specifications in a more standard manner, so that we didn't have to make adjustments for every single type of resource use measures.

And the measure modules, as I said, there were five. And within each of those five domains, so to say, or modules, together they built the measure as a whole.

So one thing to note as we go through the measure modules themselves, and you would -- you will have seen these on

previous documentations, is that last year, in working with the steering committee, they determined that, for the purposes of widespread implementation, that measure details for some of the modules could be submitted by the measure developers as guidelines. So guidelines being where we allow for some flexibility. And you -- as we go through them, you'll see where they are. And others must be specifications. And specifications mean there is no flexibility for the users. And they -- in order for them to state that they are using an NQF endorsed measure, they must follow the specifications to the letter, okay?

So the data protocol steps typically is something that NQF doesn't gather for quality measures. But it was determined by the steering committee that it was still very important, given the newness of the resource use measures, to, at minimum, have the measure developers submit some guidelines

around that. Or if they really felt that these needed to be set in stone for their particular measurement approach, that they would submit them as specifications.

And in the submission document that you received, the evaluation forms, you'll see, it will either say -- and this is the measure developer telling us whether they're submitting them as guidelines or specifications. I believe most of them submitted them as guidelines.

The clinical logic measure module, completely specifications. And the clinical logic components are the -- all the clinical steps that build, you know, the clinical, homogenous, sometimes not as homogenous, populations in which the resource use is being compared to.

And then we have the construction logic module, also completely specifications. They should be followed to the letter. And they include the steps beyond the clinical

logic. They would be, for example, while they may be related to the clinical area, they are not the underpinnings. They would be, for example, stop and end date, so 30 days after measure resource use or only include ages 18 to 55, et cetera. There are those kinds of particular algorithms.

Adjustments for comparability, again, specifications. They should be followed to the letter. They include the risk adjustment approach, any stratification. It's important to note for resource use the stratification could be for acknowledging socioeconomic differences, for example, as we discussed before.

Stratification could also be an approach that the measure developer wants people to use to make it more actionable. So let's say they're looking at a patient population of diabetes that includes Type I and Type II. And it's quite large. They may specify that those two populations be

stratified when reporting out so that those who are being measured can see the difference between their Type I and Type II resource use, as a very generic example.

And then in thinking about how the measures are reported, also an area in which NQF doesn't typically endorse for quality measures, the steering committee felt that it was important for the measure developers to think more about this and, at minimum, provide well thought-out guidance to users, but still allow for some flexibility. Because depending on the user and the perspective, they may need to adjust how they report that information out.

Or they can opt to make its specifications set in stone, you know, to the letter. So again, it's that data protocol, which includes data cleaning steps, where we allow for guidance or specifications.

And then the last module, reporting, where we allow for guidelines or

specifications.

I see some -- are there any questions? I see -- yes.

MS. CLARK: Well, I guess I do have a question on the last one, the reporting guidelines where -- how is a decision made to make that an either/or? Because it seems like -- is one of the goals to compare, be able to compare reports or data output across plans or entities that implement this? And if it is, then there needs to be something to -- you know, similar reporting, correct?

I mean, if we're saying that the construction logic and the clinical logic has to be specified, it just seems like the reporting would also need to be consistent across entities that are implementing this.

MS. TURBYVILLE: That's a great question. And this was a lengthy conversation amongst the steering committee members as well as, you know, acknowledging that NQF doesn't typically get into how measures are reported.

We endorse the specifications.

So that aside, kind of in the backdrop, the steering committee had a lengthy conversation and acknowledged that different users -- and again, we need to maybe think about who the intended users of these measures are. Is it an individual physician or is it a larger body that is then going to profile clinical sites or physicians?

But the reason why they've decided they need flexibilities, for example -- and some of them are users of measures, one user may be really interested in comparing provider organizations. And if they want to use an NQF endorsed measure that only provides specifications on how to profile individual physicians, they would not have the benefit of using an endorsed measure.

Others would need to profile physicians within an ACO. Others would need to profile health plans, as you mentioned. And so, because in this type of component

you're actually talking about what is the peer group, if they identify only one peer group in the specification, it really limits the ability for the measure to be implemented in other peer groups.

So while we want some well thought-out guidance on how you might identify a peer group, we -- the steering committee wanted us to acknowledge that there is a huge number of peer groups that may benefit from the use of an endorsed measure focusing -- you know, making sure that clinical and construction logic is valid.

Yes?

CO-CHAIR CURTIS: And just to follow up on that, I think, from my recollection, the steering committee -- that's how I was bringing it up was that these are not the end unto themselves. This is a step on the way to value.

And so for considering it from that perspective, I think it's more important

that the specifications and validity and reliability are intact, and less so that the way that it's going to be reported is important. Because this is just going to be the denominator for value or -- depending on how you calculate it. But -- so I think that's why this is a gray area within it that's not going to be as set in stone as we evaluate the measures.

That's my two-cents.

MS. CLARK: Just another comment, then. I mean, you're commenting about different people, the reports could be for different types of entities, whether it's a physician or hospital, or whatever.

I guess I'm just curious though, because most of the measures, or at least the ones I looked at, all used administrative claims data. A physician's not going to have access to the broad -- you know, the claims data. It's the payer that's going to have access to that. So how -- how would -- I

mean, I could see if a physician wants, you know, a specific type of report, but they're not going to be the ones that are having access to this information to implement, nor would they --

MS. TURBYVILLE: No, I think that's right. So payers are probably some of the more primary intended users, you know, coalitions, community efforts. You know, it's going to be situations in which they have aggregated administrative data that include enough patients or members or populations to actually support measuring resource use, so that's right.

And depending on how much data they have, they may not even have enough to get down to an individual physician level, right? So you're absolutely right. The users in many respects, and I would actually wouldn't mind the TAP exploring this further, I think are probably limited right off the bat. Initially, you need to estimate your

comparative results.

So did anyone want to add to that?

DR. PALESTRANT: I -- can you hear me?

MS. TURBYVILLE: Yes, is that David?

DR. PALESTRANT: Yes, can you hear me?

MS. CLARK: Can you turn that up?

DR. PALESTRANT: Can you hear me?

MS. CLARK: No.

DR. PALESTRANT: I'll try to speak louder. Can you hear me now?

MS. TURBYVILLE: Not really.

DR. PALESTRANT: Okay, I'll try to call back. Maybe it's a bad line.

MS. TURBYVILLE: Now we can hear you a little bit better.

DR. PALESTRANT: Can you hear me now?

MS. TURBYVILLE: Go ahead, David.

DR. PALESTRANT: Okay. Can you

hear me?

MS. TURBYVILLE: Yes.

DR. PALESTRANT: I mean, I have some issues that are sort of across -- I think it sounds like many of the people on the committee are thinking the same things. These are issues that go across the board throughout the measures.

We are being asked to individually endorse each of these measures. And some of the things that come to mind straight off were these are going to be very important metrics as we go forward with health care reform. Yet there is huge uncertainty about this data. A lot of the measures, from what I can see, have not really been validated in any great extent, nor do we even know that by measuring this, and therefore changing the way we would practice medicine to reach this score, will actually affect the cost of care.

So you know, once it has an endorsement, my concern is that it's been out

there with an endorsement, what's the mechanism to sort of ensure that this is not -- - that if this doesn't prove to be valid or actually have the effect on value of health care, what's the next mechanism then to sort of retract these measures?

MS. TURBYVILLE: Thank you, David. That's a very great question.

At minimum, these measures would come up for maintenance endorsement review within three years. So -- and that's standard for all NQF endorsed measures. So once a measure goes through an initial endorsement, and even after they go through a maintenance review, there -- we reconvene the steering committees and expert panels, depending on what the effort needs, to review the measures.

Once again, gather -- you know, we expect the measure developers to submit more information about how they're being used, and you know, there may be more questions about, from the experts, whether there are some

inconsequential consequences or are they monitoring them, et cetera?

So within three years at minimum, these measures would go through that maintenance review which, in many respects, is very similar to the initial. It's not a cursory effort, it's really a revisiting of the measures looking -- allowing for measure developers to submit new, hopefully improved measures, et cetera.

CO-CHAIR CURTIS: But I guess the concern is really how are these measures going to be used, once they have that seal of approval? And you know, I think we struggled with this at the steering committee level as to how specific we could get with the endorsement. And I think we've talked about potentially the NQF really only evaluates measures that -- or endorses measures that are intended for public reporting. Yet some of these measures may not be suitable for public reporting, in the absence of a link to

quality.

And so I think it's worth maybe taking that back up to the steering committee level to see if, in fact, we should have endorsed for some purposes but not for all purposes. On the other hand, that doesn't really -- all the measures that NQF approves have the potential to be misapplied in the population, so I don't think this is different from any of those.

DR. BURSTIN: I'll just add in that, in general, NQF does not specifically endorse measures for specific purposes. While there is now a new partnership that NQF had formed at NQF that I think is being called the Measures Application Partnership, which is now in the process of working to develop criteria to select measure for particular uses.

In this instance, though, you should assume that any measure that goes forward, that you have been recommended, goes through the whole process, commenting, voting,

approved by the board, et cetera, is deemed appropriate for a multitude of accountability uses. And so I think you need to consider that in your deliberations quite intentionally.

But again, you know, consider the fact that a good number of the folks who want these measures the most are actually community alliances and groups like that who have absolutely no cost information at the current time. So it's not clear that data will always flow down to the physician level or the clinician level. It may wind up being at a higher level of aggregation.

DR. WEINTRAUB: Well, suppose we feel that a measure is appropriate for some uses but not for others. So suppose we feel a measure of resource use is appropriate at the hospital system level but inappropriate at the level of the physician. Can we make it clear that we think it's -- or is that beyond our scope?

MS. TURBYVILLE: Yes. And there is an actual component of the measure submission, and so the specifications, that ask the measure developer to cite what the appropriate unit of analysis is and certain the TAP and the steering committee can say, we think the unit analysis should not include the physicians, but it's okay to include -- et cetera. So that would be something that you would -- we would facilitate that through to the measure developers.

DR. MARWICK: Could I seek some clarification between what we're discussing now and the timeline? I mean, I could see a number of these as being works in progress where eventually a useful measure would be constructed. But at the moment, in my opinion, the measure is not useful.

I have the impression that you're committed to approving at least some of these in the course of this year. That fills me with some disquiet, I'd have to say.

MS. TURBYVILLE: So as is true with NQF endorsement process for this, the expectation is that the measures that are coming are fully specified and have been tested adequately. You're right, the timeline does not allow the developers a lot of latitude to go back and especially respecify a measure, because they would have to then figure out how to retest it within the timeline.

Often we can expect that they can make some adjustments, if it becomes critical in order for the measure to move forward. But there -- and we rely on the measure developer to state what they can and cannot accommodate, if there are requests for them to make adjustments. But you're absolutely right, these should be -- for today, what we're looking are their full specifications, knowing that there may be some back and forth, but their ability to make drastic change is very likely to be limited.

Again, that would be up to them to react to, but you can imagine them having to change a specification, get it through their experts, test it, et cetera, would be certainly a time crunch.

DR. MARWICK: So you're still committed to approving some, even if they're not satisfactory, correct?

DR. BURSTIN: No. We will only put forward the measures that this group, the steering committee, the membership, the broader population, agrees are acceptable. And it may very well be there will be very few measures at the end of this process, and that's okay. It's our first foray into resource use.

I think that we will see the measures continue to get refined. We will likely see measures coming forward in the future that actually do combine cost with quality. I think the issue is at this point, we know we -- we just knew we needed to get

off the dime and start somewhere. And there's obviously a great deal of interest in having these measures out there.

I will tell you, at the population level in particular, a lot of these communities that are working in community alliances have nothing to assess where they are. So for a lot of folks, there is a sense that we should, at least, start, see what's out there, see what's doable, find out -- a lot of what emerges out of these projects actually is suggestions for improvement, suggestions for additional work to be done.

So I see this as a first step in a path.

DR. MARWICK: Sometimes bad information is worse than no information.

DR. PALESTRANT: I think without linking this to quality, you're -- you know, I think it's an admirable first step. But without having a link to quality, what are we measuring? I mean, that's essentially the big

question. Sure, if we give no care it costs less. But that's not our metric. Our metric really is outcome. And that is an essential model of value.

And if we're going to be measuring value at some other criteria, using different databases, how can we be sure that these databases and costs are actually the same databases that are going to be used for measuring outcome?

DR. BURSTIN: This is Helen Burstin. I'll try that.

In general, when we've talked about this to date, there has been an expectation that we see these measures as building blocks. Not ones to be used on their own, but ones that we at least needed to begin to understand their construction, the issues involved in them. The committee spent a great deal of time thinking through what -- how our current evaluation criteria, which are really designed for quality measures, work or could

be adapted to make it fit for research use measures.

Again, it's really just this first foray into it. I think a lot of the general outcome measures we have ultimately could be knitted together with some of these resource use measures, the more longitudinal measures, for example. You know, a diabetes measure over a year, for example, some of the logical outcomes are easier than I think perhaps some of the other conditions.

Again, we expect this to be a journey, and it's just our first step on it. Nothing will come out of this that people don't agree is ready for primetime.

CO-CHAIR ROSENZWEIG: Could you just describe the implications of NQF endorsement, what it exactly means in terms of if a plan -- does it mean that a plan -- a plan is free to use this if it doesn't have NQF endorsement? Or I remember the government organizations need NQF endorsement?

DR. BURSTIN: So the basic guidance for an NQF endorsement is NQF is a standard-setting organization under the National Technology Transfer and Advancement Act. So essentially as a national standard-setting organization, it makes us the measures of first choice.

Essentially, when the federal government -- and in particular it mainly is applicable to government, although increasingly as we're seeing harmonization across the broader quality enterprise, a lot of plans are also looking towards NQF for endorsement. There is an expectation that they will look for NQF endorsed measures first and use only -- and use non-NQF enforced measures if NQF measures are not available, or if they could justify the use of a non-NQF endorsed measure.

Actually, as part of ACA, there needs to be a posting in the Federal Register when the federal government chooses to use a

non-NQF endorsed measure and explain the logic. So it does have a fair amount of heft there.

But again, keep in mind plans can use anything they want, and they already do use a good number of the commercial groupers here, and the costs associated with them.

DR. WEINTRAUB: Another thing to give a little perspective from someone who's been involved in cost effectiveness analysis now for over 20 years. This has been said about cost effectiveness analysis, if you've seen one cost effectiveness analysis, you've seen one cost effectiveness analysis.

And a standardized approach to costing would actually be very, very, very helpful, but as Tom says, bad information is sometimes worse than no information. So ultimately this has got to be gotten right, because these measures will be -- will be used really extensively, and I think far beyond what we're even dreaming about here today.

Especially if they're really -- if they can be validated well. She said it's going to be hard to go back and respecify. Well people may have to go back and respecify, but -- in what I've read, I'm less concerned about the specifications than the testing and the validity, which I think for measures like this, and as you say it's your first foray, this has really got to be very, very extensive.

Now the ones that I read were not the ones that are being used, so I'm looking at ones that are earlier in the process. But I think that the validations, so that we can believe it and so that others can believe that these measures are really measuring cost, has got to be pretty rock-solid. And that's for someone who's spent years measuring cost, I can tell you it is extraordinarily difficult to do.

MS. TURBYVILLE: Great. So I think at this time we're ready to hand it over

to Jeptha and Jamie. Though I think we also -- pardon me, before we do that, I think we want to talk about the voting, right?

So if you look at your cheat sheet that looks like this. And then Sarah's going to talk about how the scaling works, which has another cheat sheet on the second page, the smaller table that looks like this.

Just for your reference -- yes, that second page -- it outlines the tasking -- the testing task force guidance on what high, moderate, low, indicate for reliability and validity. So you'll see the first column is a validity rating. And you'll see a "high," what you want to expect to see: evidence of reliability and validity. And then that would be moderate to high. And whether or not it passes scientific acceptability for that particular component would be yes.

You can see for the low ratings, that indicates a no for passing scientific acceptability based on your ratings. So I

just want to make sure that we're all at least on the same page of what high, moderate, low means. And as we all get fatigued throughout the day, you have this to refer to if need be.

Okay? All right, great.

CO-CHAIR CURTIS: So we're 25, 30 minutes behind schedule already, which is not bad for the first hour and a half. And I think, per the agenda, we're supposed to start off with Measure 1571, which I'm the lead reviewer, and I believe Mary Ann is the secondary reviewer.

I think, in terms of process, you know, we have -- how do we get your attention? You know, you can flag your keychain or you can fold up your tent and we'll try and get everybody involved in this to have comments.

I think the one question I have is, do we have -- in previous efforts we've had flash drives that have had all the measure specifications so that everyone can have them in front of them. Okay, I think it --

MS. WILBON: We do have flash drives, and we also have, I think in the packet that we sent out, there's links -- if you guys have access to the internet, there's links to each of the measure packets on that. So you can -- either one, whichever is easiest for you. If you're having trouble accessing the internet, we definitely have flash drives we can pass around for you to download, and we'll try to bring it up on the screen as well. So just raise your hand if you --

CO-CHAIR CURTIS: I think this will be important since, you know, only two people have seen the -- really gone through it in great detail, and make sure we're talking about the same thing.

But I think we can try it on the screen --

MS. WILBON: Yeah.

CO-CHAIR CURTIS: -- but we may have to, in the next set of measures, go so that we're all looking at the individual

screens.

CO-CHAIR ROSENZWEIG: What's the URL for the internet?

MS. WILBON: For the measures?

CO-CHAIR ROSENZWEIG: Yes, I don't have all the measures. I just have the ones that were sent to me.

MS. WILBON: Oh, okay. Why don't I give you the thumb drive.

CO-CHAIR ROSENZWEIG: Oh, the --

CO-CHAIR CURTIS: But for everybody.

MS. WILBON: The links are actually in a document that we sent out in the pre-meeting -- the pre-meeting email. So if you don't have access to that then we will pass around the thumb drive. We have a few thumb drives we can circulate, if you need access to the documents.

CO-CHAIR CURTIS: I think, since we expect it's going to take about an hour to go through this, why don't we take a five-

minute break for restrooms and come back quickly then.

(Whereupon, the above-entitled matter went off the record at 10:23 a.m., and reconvened at 10:32 a.m.)

CO-CHAIR CURTIS: So I think maybe we should get going so we don't fall even further, which is going to happen. But I think everyone's back at the table now.

So we are scheduled to go through 1571, which is an ABMS measure, acute myocardial infarction episode of care for post-acute period, parentheses, days 31 to 365.

As I mentioned, I'm the primary reviewer and Mary Ann is the secondary reviewer on it. And it should be -- for the people who have it on their computer and can bring it up, and we'll bring it up although it's a little hard to read. I don't know if we can make that slightly bigger.

So I'm going to walk through the

measure fairly linearly, trying to respect the time-frame. We sort of set aside an hour to go through this and obviously there will be a lot of cross-cutting issues with the other six ABMS proposed measures.

So the overarching picture here is that it's -- they're trying to characterize resource use and costs associated with AMI care during the post-acute period. And they're defining that post-acute period as 31 to 365 following an index AMI event. And I'll -- yes, resource attribution is at the level of the individual provider, the other key. It's specified using exclusively administrative data. Although they do say other, and I'm not sure what other means. But I don't see it anywhere else in the application.

It's I think a little odd to consider this in isolation because this really is a paired measure with the acute episode, the first 30 days of the AMI care, as well as

-- and this would be the paired measure -- with the follow-up care. And I think probably the issues raised on this application are going to be almost identical to those of the earlier phase.

The measure developer I believe is not on the phone.

MS. WILBON: We can double-check. Kevin, are you on the phone?

DR. WEISS: Kevin Weiss is here. We've got Todd Lee.

MS. WILBON: Todd Lee as well?

DR. WEISS: And we have Robin Wagner.

MS. WILBON: And you have Robin, okay. Do you know Kevin, we're having a hard time hearing people on the phone, so we ask that you speak up. They have adjusted the volume, but it's still not as loud as we would like.

DR. WEISS: Okay. Well, mindful of that, is this a little bit more helpful?

MS. WILBON: Yeah, that's much better.

DR. WEISS: And also, with apologies, I'm going to be boarding a international trip shortly, but I'll be here for a while.

MS. TURBYVILLE: Okay. So do you want to take some time, Kevin, to introduce the measure and the approach that ABMS-REF already took in developing the measures that they're going to review today.

DR. WEISS: It would be a pleasure. And I want to thank the committee for the opportunity and NQF for sponsoring this review.

Let me start by saying that I have to just put a small note that this is technically not ABMS, the American Board of Medical Specialties, but it's the American Board of Specialties Research and Education Foundation, the REF. That's not an insignificant difference because I think one

of the issues, because of ABMS is the question, because ABMS is a standard-setting organization for physician certification, is this something that is directly going to be moved directly into the certification MOC program. And currently that is not slated as these are new measures, and it's the first time for the ABMS-REF to work in this environment.

Notwithstanding, what I'd like to say is, the way that these measures were developed were based upon the interest of getting the provider community, principally, but not exclusively, the physician community to fully develop and endorse a set of resource use measures that they themselves felt had strong face validity. And then build from that face validity into all the other constructs of a good measure specification.

They were built with a very keen interest, and that is that they were viewed to be eventually paired with quality metrics so

that resource use and quality could be brought together for the concept of value or efficiency. So these were constructed not on -- based upon resource use flow of data but rather based upon the literature surrounding the quality of care of episodes.

So the episodes were derived from a clinical perspective. The timing of the episodes were derived from the perception that, of experts in the field, as to what the literature would suggest that the episode that is being thought of in usual clinical care. And so you'll see them constructed that way. The AMI measure was constructed that way and it's broken into two pieces for that following reason, as well as attribution.

As far as attribution, we looked at all the measures to try and get to the level of individual physician attribution, if it was felt appropriate. But we were very clear with the measure work groups that developed these measures that the providers

and our technical advisory committee who oversaw the process, to say that if we felt that it was inappropriate to do physician attribution, that we would propose higher aggregation where it was appropriate. And the paired measures you'll be looking at first from us, first this longitudinal or follow-up care for AMI and its paired measure of acute AMI are very different because the 30-day measure, which you'll be reviewing, I believe, tomorrow now, is not being viewed as an attribution to a physician. It's more of a system measure, and that will be explained more tomorrow if you're -- if you would like us to.

And this measure, after 30 days, it was felt that care does -- of a patient does move into individual physician care with actual patient or consumer choice. And so that's where one could begin to try and seek an attribution at individual physician level.

So with that in mind, we recognize

-- and you'll be seeing a common theme, and we heard it even in your discussions, about the nature of where we have developed these measures, in terms of advanced testing. We recognize, as you do, that ideally these measures would have a community-based field testing, or some type of field testing. We are in the process of doing field testing. We have several communities who are getting our data and are beginning to evaluate that. We've worked with a couple of other data sets, and those we can talk about as well.

However, we do feel these measures are fully specified and have gone through a rigorous review process of the specification and the initial validity and reliability testing, and felt confident that they met those criteria that NQF presented. We hope that the pretty clear face -- face valid issue of lack of community testing is not going to be the key issue that holds these back, but we would respect wherever the committee goes on

that, of course. But as long as you know that we feel that the next step is field testing, and we're actively engaged in that process.

I'll stop here. That was just a very brief overview, but in case you have any general questions you'd like to ask before you get into more details on the AMI measure specifically.

CO-CHAIR CURTIS: So it doesn't seem like there are any specific questions yet. But I know we will be -- or I will be addressing questions to you, if not other members of the committee, as we go along.

So I think with that, let's leap into the specific criteria that we're addressing today. And the first is the issues of importance to measure and report. And I think that probably for most of these measures, the important issue is not going to be the major one that we're evaluating. And so I don't want to spend too much time on that, although I'm certainly willing to open

it up for discussion if people think differently.

But of course, this is a measure of AMI care and resource use associated with the care of AMI patients, which are a high-risk, vulnerable population who consume a lot of resources and are vulnerable to lots of adverse outcomes. So they cite the usual panoply of information suggesting that this is an important population, and I agree.

And in addition, the -- so I think, is that 1.A criteria? So I don't know, do you want to do that as we go along or finish importance and then vote? Okay.

In addition, 1.B is the opportunity for improvement in disparities. And again, they do a fairly light literature review citing variations in the care and outcomes of this patient population. But again, I don't think it's one that is terribly contentious.

I will say, though, in specific to

this measure, is that I don't see any evidence that they're providing -- I'm sorry, they don't providing any evidence that there are, in fact, variations in this post-acute care time-frame. And while I know, as a clinician, that they do exist in terms of the intensity of monitoring and likelihood to treat medically versus refer for surgery or percutaneous interventions, I know that that's there, but I don't see any empiric evidence to back that up, which I thought was a little bit of a limitation of their importance.

Mary Ann?

MS. CLARK: I just want to echo that, because that's what I found as well. It seems like this is definitely an important area, this post-acute care period, but there were no citations on variation and resource use across that time period that I saw.

CO-CHAIR ROSENZWEIG: Citations of variation were actually in the pre-30-day period, looking at the references. I don't

know, I didn't go through all of them in detail, but several of them specifically seemed to refer to acute -- sort of not the post-acute care, but the more closer to acute care.

CO-CHAIR CURTIS: Agreed. That may just be the limitations in the literature, which really have not traditionally focused on this post-acute care, or broken that out. You know, there have been certainly long-term ones, like three, five-year outcome studies of AMI populations. But breaking this part of it out, there just may not be literature there.

I don't know if the measure developer wants to comment on that?

DR. MARWICK: Could I ask some guidance about the voting here?

So in relation to our voting about this, are we assessing it in terms of the importance that we perceive or the degree to which that measure has been addressed in the document? What I have in mind is that there

may very well be measures that we understand are very important but in fact have not been well defended in the document. And if they are -- if our support is important in terms of getting that material released, it could end up being sort of embarrassing that this is not a well-prepared document that's finally approved.

DR. WEISS: Kevin Weiss. Did you want me to address that question? You asked if the measure developer would like --

CO-CHAIR CURTIS: That would be great.

DR. WEISS: We, along with you, recognize that there is no literature that really speaks to this period. When we brought our panel of experts together we said to them that in many cases there is a very weak bit of literature that defines variability in resource use. We wanted to use their clinical experience collectively to sort of clearly identify what they saw in practice as an area

for potential resource variability.

What they specified here was, although there is no resource -- not a -- there was not really a body of literature, that there was a high degree of perceived use of intervention in a time period where the guidelines do not speak to the need for these interventions. Specifically, extra stress testing and extra invasive percutaneous treatment. So the lack of literature does not bespeak the missing of that literature, at least to our knowledge.

DR. BURSTIN: And in terms of the importance question that was raised, we would very much like you to stay grounded in the criteria. So your voting on the criteria should reflect directly the questions the criteria asks you as what you see in that application, not a broader context.

DR. WEINTRAUB: I might just comment in general on the issue of disparities and resource use. I've participated in

literature on disparities and participated for years in literature on costing for years. And these are literatures that almost don't overlap.

You know, I think that that's one that we're going to have to, to some extent, give them a bye on, because that literature isn't going to all of a sudden come into existence in the next couple of years in any robust way.

DR. PALESTRANT: I'm not sure if anybody else can speak further --

CO-CHAIR CURTIS: Yes? What was the question? I'm sorry.

DR. PALESTRANT: Oh, sorry, can you hear me?

CO-CHAIR CURTIS: Yes.

DR. PALESTRANT: Yes, this is Dr. Palestrant, can you hear me?

MS. TURBYVILLE: We can hear you really well now.

DR. PALESTRANT: Okay. The -- if

anybody is in the room who is more of an expert in cost analysis and resource use, I mean, I know that there's some discrepancy regarding the document's data. Yet all of these -- all of these measures are really referring back to the document data and sort of citing it as absolute. Aan anybody comment on what the sort of -- the thinking right now is, in terms of that data?

DR. WEINTRAUB: Yes, I can comment on it.

From the point of view of variability, especially regional variability and resource use, there the data are -- as opposed to healthcare disparities, on gender/age/race -- the data on geographic variability, much of which comes from data, is pretty exquisite. There we know there certainly is variability in resource use almost anywhere you look.

DR. PALESTRANT: Right. I know from personal experience that, in some of this

data, by simply changing or increasing the coding that one is doing in institutions, one can affect that data. And not really changing anything in terms of your resource use or anything else that you're doing, but simply having people document more clearly what the diagnostic codes are, changes our outcomes in terms of observed/expected.

This is -- you know, I guess it gets back to this point, what is a real valid measure on an institutional level?

CO-CHAIR CURTIS: So I think that's a very important point, and we'll probably come back to that as we get to the measure specifications. But I'm going to table it for the discussion of the importance.

Okay. So then -- and then we didn't get to, but Bill alluded to, the issues of the disparities. And certainly there are -- there is documentation of disparities in the quality of care delivered by age, race, gender and socioeconomic status. Again, I don't

think that there's a lot of information about disparities in terms of resource use or cost.

And again, we'll get -- I think come back to this in the review. But I think it's something that's important to consider in all these measures is, in fact, do we need to consider disparities in resource use? And is that something that's important for measure stratification? But certainly we'll come back as we go through the measure criteria.

So moving to the opportunity to -- sorry, not the opportunity to improve but the -- for measure intent, I think it's fairly straightforward. I'll just read what the measure developer wrote. "The intent of the measure is to provide an estimate of the overall resource use associated with an AMI care and identify components of care that are most associated with high costs. Providers can be compared in terms of their relative resource use, compared to their peers and reasons for differences in cost can be

identified. Ultimately the measure needs to be combined with quality for a measurement of efficiency of care."

So I think that was a pretty nice description of the intent of the measure. And I didn't have any particular criticisms of that.

And then regarding -- sorry, is that 1.D? 1.C, sorry.

1.D is -- I've lost my train.

MS. TURBYVILLE: 1.D is about the resource use service categories being consistent with the intent of the measure.

CO-CHAIR CURTIS: Right. And so we'll get into that, I think, as we get to the specifications. But my overall impression was that they were, in fact, consistent with the intent of the measure. So I'll leave that open for anyone else to comment on, Mary Ann specifically.

MS. CLARK: So yeah, the different categories of resource use that they had, and

this was consistent across all those measures that I was evaluating. I think there was a little bit of a lack of clarity on how those were being defined. If you can -- I don't know if you can find those in the document, the resource use groupings, categories.

MS. TURBYVILLE: For the categories?

MS. CLARK: Yeah.

MS. TURBYVILLE: Yeah.

MS. CLARK: There you go. So there was just a little bit of clarity that I think needed to be provided here. For example, it's broken out into broad categories of inpatient versus ambulatory services -- whoops.

And then within each of those broad categories, it's further broken out into several different components.

Well, I guess my question is, how is -- how are physician services captured in this? You know, there's going to be a

physician component to any inpatient or outpatient service that's being provided. So I'm just wondering where that's actually captured? For example, you have inpatient facility services under inpatient. And then procedures and surgeries. I mean, I guess I'm just wanting a little more clarity on how those categories are defined.

Same thing for outpatient. You have outpatient facility services and then procedures and surgeries. Is that what the distinction is? I'm not quite sure how those are defined.

MS. TURBYVILLE: Before I hand it over to Jeptha, I do want to note that these are check boxes that NQF put out there for them to check which services. So -- but I don't -- so they could check another if we didn't encompass the universe.

MS. CLARK: Uh-huh.

MS. TURBYVILLE: But Jeptha, I didn't know if you had a response to the

details of the questions, but there was approximately how that works.

MS. CLARK: Oh, yeah.

CO-CHAIR CURTIS: Yeah. I mean, I guess I just didn't recognize that there were any missing domains here. I mean, I think physician services would be under the evaluation and management of both the inpatient and the outpatient, and that the procedures and surgeries would cover the actual interventions, per se.

So I thought it was comprehensive -- well, at least I couldn't think of any domains that were missing.

MS. WILBON: And also just to note, in the specifications section, there is a question that asked them to actually define. These are the resource use service categories, so I think some of your questions may have been addressed, or should have been addressed in that section. And it may be a little more clear when we get down to that in the

specifications.

CO-CHAIR CURTIS: So in the interest of moving along, you know, I think we could probably vote on the importance criteria then. And I guess you should take out your keychains and remind us what's high and what's low?

MS. TURBYVILLE: Yes.

MS. WILBON: So on the monitor here over to my left, your -- well, some of your right -- is the -- we'll be pulling each of the subcriteria on the screen and you can see the definitions for high, medium and low. Once we hit the timer, you'll have -- once we hit "start" you'll have 60 seconds to enter your vote.

If you voted and then you want to change it, there's a hazard key on there, like a triangle with an exclamation point in there. Hit that button, enter your new rating and then hit "send." So if you mess up your score, hit the hazard key, enter your new

rating and then hit "send," and then it will recalculate it. And then once everyone has submitted the -- the results of the voting will show up on the screen?

MS. CLARK: How did you know if it's one, two, three?

MS. WILBON: It's high is one, moderate is two, low is three, insufficient is four and not applicable is five.

CO-CHAIR ROSENZWEIG: So we're really voting on what we believe to be --

MS. WILBON: Right, based --

CO-CHAIR ROSENZWEIG: -- the importance? So at least in theory, if we've read some of the others and they provide evidence for importance --

MS. WILBON: Yeah.

CO-CHAIR ROSENZWEIG: -- we're actually voting on what we perceive as the importance for this measure?

MS. WILBON: For this particular measure.

CO-CHAIR ROSENZWEIG: It's not how well they particularly described it?

MS. WILBON: Yeah, well, it's based on what they submitted. So I'm -- as Jeptha mentioned before, we're not implying or making any extrapolations based on what they submitted. So if they --if you felt like what they submitted may not have been incomplete and that may be the rationale that was provided by the developer, there not being evidence in the literature, whatever, however you feel that what they submitted, based on what the subcriteria is, that's what you're rating should be based on.

CO-CHAIR ROSENZWEIG: So really, we're voting on how they submitted it as opposed to what the truth -- whether or not we view this -- the importance to be high or low?

MS. WILBON: Right.

CO-CHAIR ROSENZWEIG: Okay, I'm sorry, I --

MS. WILBON: It's based on what

they submitted, yes.

CO-CHAIR ROSENZWEIG: Okay.

MS. CLARK: Compared to the subcriteria?

MS. WILBON: Right. Is everyone clear on that?

CO-CHAIR CURTIS: It's important to verify.

DR. HWONG: I guess what could be interesting is, I mean, what you could is, you know, we're talking about diabetes and we think they are -- in terms of what we understand about diabetes, variation care, potential disparities in the management of that, we're really -- you know, for one measure may rank the importance as high, even though, you know, let's say they're both measures that look at chronic care of diabetes.

I'm just saying that like somehow you'll have some sort of discrepancies in sort of high versus moderate versus low, even

though it's really talking about the importance of this condition in terms of management of chronic care.

And if that's the case, that's okay, we'll just -- you know --

MS. TURBYVILLE: Right. So I think, though, and that comes out in the justifications, when we ask all of you to justify the high, medium, low, so that either the developer, especially for importance, may say, okay, we'll submit more evidence. Or if it's really an issue that cannot be -- it's a hurdle that they can't meet -- Helen, I don't know if you had anything to --

But it's really important to stay focused on the subcriteria as has been determined by the steering committee and kind of public scrutiny and based and rooted in criteria that NQF has used for a long time.

So the first one is the measure addressing an important focus area. And by demonstrating that either it hits on one of

the DHS national priorities, or it's a high impact area of health care, which -- large numbers, et cetera.

CO-CHAIR CURTIS: So I take it I'm supposed to go through the voting. So then on -- starting on subcriteria 1.A, whether the measure addresses a specific national health goal priority, or the demonstrated high-impact of health care affecting large numbers, et cetera.

So submit your vote.

CO-CHAIR ROSENZWEIG: We can't vote twice?

CO-CHAIR CURTIS: The number and then "send," correct?

CO-CHAIR ROSENZWEIG: Oh, the number and then "send." Oh, send?

MS. WILBON: No.

CO-CHAIR CURTIS: Oh, just the number. Got it.

CO-CHAIR ROSENZWEIG: Say that again?

MS. WILBON: So it actually tells us how many responses.

CO-CHAIR CURTIS: So unanimous for high for 1.B demonstration of resource use for cost problems and opportunity for improvement. And go ahead and send.

And let me just ask while we're voting, I mean, I think as the reviewer, no one has the benefit of my pre-review because I failed to submit it. So I will feel free to give you that as I go along.

MS. WILBON: And Dr. Palestrant, if you could be using the document I sent you to enter your ratings, that would be great.

DR. PALESTRANT: Oh, okay. Just email the rating to you?

MS. WILBON: Yeah, you can just collect them through the course of the day and send them to me at the end of the day.

DR. PALESTRANT: Okay.

MS. WILBON: All right. Thanks.

CO-CHAIR CURTIS: Five moderates,

two highs and one insufficient.

For 1.C, the purpose of the objective resource use measure and the construct for resource and costs are clearly described in that measure of intent. And I rated that as high.

Seven highs and one moderate.

1-D is one of the ones we might have to come back to, upon further reflection, but at least we'll get a preliminary vote at this point. The resource use service categories that are included in the resource measure are consistent with and represented above the conceptual construct represented by the measure. And I rated this as high.

And more evenly split between high and moderate.

Okay. So moving on to the heart of the application, which is the measure specifications, and evaluating it using 2.A.1, 2.B.1, et cetera, we're going to walk you through -- it's a little hard to walk through

all of them, but I think we kind of have to.

You might want to keep in mind what I'm considering the other cheat sheet, which is the submission items that are affiliated with each individual criteria, which is, I guess, page 2 of the evaluating resource measures. And so for 2.A.1, for instance, go over the general approach for resource use measures, et cetera, et cetera.

My feedback to the steering committee and NQF would be that there are way too many sub -- or I will call them microcriteria as opposed to sub-subcriteria, to be evaluated within one larger subcriteria. But I'll leave that open for other people's feedback as well.

So the general approach, as the developer discussed, was that they started with a work group working in conjunction with analysts to derive what I think are fairly clinically sensible approaches to defining a coherent population and coherent outcome, and

a reasonable risk adjustment methodology. The talk about it's an iterative process, that they went back and forth refining the -- and specifically the outcomes, not so much the population that's being used. And they provide I think fairly ample supporting information as to the data dictionaries, et cetera.

So with regards to the specific resource use measure, this is a standardized cost measure using administrative data. The target population is patients admitted with a index principal discharge diagnosis of acute AMI. But the outcome period, as mentioned before, is 31 to 365 days. So the post-acute phase care of this patient population.

Regarding F-6-1 and 6-2, the data preparation inclusion, which I believe are on pages 10 and 11 of the PDF, they make guidelines, as opposed to specifications, as to how to handle the cleaning process. But mainly defer to the individual providers as

having the internal expertise and allowing leeway for specific handling of data.

They do strongly recommend that missing data not be included and that no approaches for imputation be utilized, which I think seems reasonable, and I think is consistent, at least across the two developers that I reviewed.

CO-CHAIR ROSENZWEIG: Can I just ask you, is the definition of 31 days after acute MI a recognized interval in the cardiology community that defines post-acute? I mean, it's not after discharge from the hospital or something like that?

CO-CHAIR CURTIS: Right. So the measure developer doesn't specify why that interval was chosen or selected. I infer or assume -- there is literature supporting 30 days as sort of a reasonable timeframe for the acute care. So if you think about the publicly reported measures for AMI mortality and readmission, they utilize a 30-day post-

discharge as their episode.

And so this is consistent with that and I would imagine they selected this particular timeframe, breaking it out of 30 and 31 to 365 as being one step in the harmonization of resource use measure with a quality measure. But again, I could ask the measure developer to commend specifically on that.

DR. WEISS: This is Kevin. Can you hear me okay? I switched to a cell phone.

But the answer is "yes," if you can hear me. It's because we first developed the acute measure, which really was based upon very, very substantive literature as well as a convention within the cardiology community. And it was recognized that there needed to be an extended period which seemed to be also based in one-year outcomes in the cardiology literature. And we started at the 31 and went to the end of one year, which would provide a nice eventual harmonization with outcomes

measures, if we went that way.

Is that -- are you able to hear me okay?

CO-CHAIR CURTIS: Thank you.

Bill?

DR. WEINTRAUB: Yeah, so it's in the clinical trial literature, a lot of analyses that are zero to 30 days and 30 days to a year. There's nothing particularly about costing in that period, but it does harmonize with other measures of outcome.

CO-CHAIR CURTIS: So with microcriteria 6.2, data inclusion criteria, again they are fairly clear that they're basing it on the finalized cohort as opposed to any preliminary cost data. So this is something that would not subsequently change. So the database is finalized and complete.

They recommend -- and we'll get into the specifications, but in order to calculate the risk-adjusted costs and utilization, they require that enrollees have

at least 24 months of continuous medical and pharmacy benefit enrollment, including both the identification year and the measurement year, or I'd say that that's recommended. But I assume it's almost a requirement.

They know, however, that the measure was tested on enrollees with at least 320 total days of coverage during each year, which I assume is a nod to the practicalities of the database that they had to develop the measure.

So any comments on that particular element?

DR. HWONG: Right. So in essence, they're defining continuous eligibility with that criteria of having at least 320 days?

CO-CHAIR CURTIS: At least pre- and post-period.

DR. HWONG: Uh-huh.

CO-CHAIR CURTIS: I think that's a function that they had two years of data -- it must have been at least two and a half years

of data to work with to derive it, because the index submissions took place, I think, between July of '06 and December of '06.

And then regarding the data exclusion criteria, 6.3 --

MS. CLARK: Just one last comment on that. It would seem like you'd need at least three years worth of data to do that type of analysis. Because if somebody had their AMI at the end of the year, you know, you're not going to have the full follow-up period, but yet you need a full year of look-back in order to assign the hierarchical condition category risk adjustment method.

So I think three years is probably the minimum.

CO-CHAIR CURTIS: I would agree with that. And if you look at the dropout from the inclusion criteria from the cohort of studies, you see that there is substantial dropoff. We'll come back to this, but I think they lost about 30 percent or more of the

population, maybe 40 percent of the population didn't have continuous enrollment in both the pharmacy and the services providers.

And does the developer want to comment on that time period for assessment, because I think that is a real reality of -- that you would require three years as specified, or up to three years, if you're using a calendar year.

DR. WEISS: I'd like to see if -- Dr. Todd Lee is with us, and maybe he could jump in?

CO-CHAIR CURTIS: Yeah.

DR. LEE: The committee is exactly right. It requires a three-year timeframe to be implemented in the way that we've specified the measure.

DR. HWONG: Can I just ask -- I'm sorry -- one other clarifying question?

So the acute, you know, MI has -- in terms of making sure that it has to occur during the measurement year. But in order to

assess the resource utilization, could you -- does this end up having variable like follow-up time to the end of the measurement year?

I'm hoping that makes sense. But if the AMI -- if your acute MI happens in December 1st and the end of your measurement year is December 31st, you have one month of follow-up to look at resources -- or rather, let me say two months, because it's a 30 to, you know, 365 --

CO-CHAIR CURTIS: Right, but the -- and I think that's why it's a three-year measure --

DR. HWONG: Yeah.

CO-CHAIR CURTIS: -- is that for that patient admitted on December 30th, they have to have one-year following. So a three-year measure.

DR. HWONG: Okay, very good. So in some ways, we could specify kind of when the event has to occur in the relative timeframe? Like -- and maybe I missed that,

if it's very specific. But it has to have -- you know, the event has to have a full 12 months, you know --

CO-CHAIR CURTIS: Right, and they do specify that you do need the continuous -- or the continuous enrollment so you have both the --

DR. HWONG: Both, okay.

CO-CHAIR CURTIS: -- upstream for risk adjustment and the downstream for accounting costs --

DR. HWONG: Okay.

CO-CHAIR CURTIS: -- and research use.

DR. HWONG: Perfect, thanks.

CO-CHAIR CURTIS: Okay. So then regarding 6.3, data exclusion criteria, it's fairly straight-forward is that they recommend eliminating all rejected and unpaid claims, which again seems consistent across developers.

They also recommend getting a --

because they're attributing to the level of the physician, they recommend generating a uniform specialty for all providers, and not utilizing claims where you cannot identify a single provider using a hierarchy that we will come back to.

And finally, converting missing and zero quantities at a minimum -- to a minimum of one to allow for pricing of these services, which I have to confess is beyond my specific level of expertise in costing, so I'll defer to the cost experts.

DR. WEINTRAUB: Can you -- well, just let's look at that again.

CO-CHAIR CURTIS: So -- sorry. To repeat it, so they're converting zero or missing quantities to a minimum value of one. It allows for pricing of these services.

Would you then clarify that -- would the developer clarify that rationale for that particular decision?

DR. WEISS: Sure. So this has to

do with just the quantity field. If the quantity field has a missing value but yet there is a submitted claim that has made it through and has a dollar value associated with it, we did not want to get rid of that information. Rather, we assigned the quantity units to one in that case, so when we're calculating our average costs, we still use the actual paid claim in that calculation.

CO-CHAIR CURTIS: And for some of your costs, I think ancillary services, when you're sort of average -- developing an average cost for a service, would that tend to lower the average cost, I assume? And number two, how frequent is that in the data set?

DR. WEISS: Unfortunately I cannot answer the second question off the top of my head. I don't know -- we don't have our programming folks on the phone to answer the frequency with which it occurs.

And I am also unsure of which direction the bias would go in. It really

depends on the amount of that claim, for which there is missing quantity information So if it's a high-dollar value claim that happens to have a missing quantity information, it could bias the average cost upward, and vice versa.

But I can look in some of our files and find -- see if I can find the answer to those things.

CO-CHAIR CURTIS: Thank you.

And then finally, regarding missing data, to reiterate that they recommend not using imputation to replace missing data. So I'm going to pause there. So that's sort of the data handling and processing part, I believe. Is there more that I'm missing? No, I believe that's it.

So next, moving on to the more clinical framework of the measure, which starts with criteria 8.2 and beyond. This is, again, the resource use from 31 to 365 days that's attributed at the level of the individual providers.

For inclusion criteria, and this is starting on page 12, I think they applied fairly straightforward, or to me what rational decisions limiting the population to 18 to 85. Now the 85 warrants the high-end exclusion, patients above 85 warrants a little bit of thought. Their rationale is that it's a different population in whom treatment decisions may be significantly different than in younger populations. And so that the resulting costs may have biases probably lower rather than higher. That seemed like a reasonable choice to me, but again could be interpreted both ways.

DR. WEINTRAUB: Well, if I was developing, I wouldn't do that. And you know, we've moved away from the idea of upper - high-end limits for clinical trials. And given that this is about acute myocardial infarction, acute myocardial infarction is common in elder, including the very -- the elderly and above 85 wouldn't be my choice.

I don't think it's wrong, it just wouldn't be my choice.

CO-CHAIR CURTIS: That's fair. But also when you look -- if you get to the exclusion criteria, the percent of the population that it applied to, I think it was less than one percent, or a very, very small number of claims. That's particular to this commercial database. Obviously if this were a CMS database, it would be a different matter altogether.

DR. WEINTRAUB: We expect that that's what's going to be most commonly applied. That's the fastest-growing portion of our population.

MS. TURBYVILLE: Just to be clear, so depending on where the measure is tested -- for example, this measure has been tested in the commercial database -- that's what we're endorsing it for use in, is the commercial population. I just wanted to make sure we're all on the same page on that.

CO-CHAIR CURTIS: Okay. So then the other specific inclusion criteria is that they have -- the index is an admission for -- with an ICD-9 at 410.XX, excluding X.2, which suggests a, in fact, an acute MI and not subsequent care of a patient with a prior MI, that they are applying it to a calendar year measurement. They have specific exclusion criteria, notably in terms of enrollment criteria and both medical and pharmacy benefits. And they do apply a requirement of a length of stay of greater than one day, which is probably -- has to do with, you know, face validity of whether or not it was actually an MI. And I think even in the days of decreasing length of stays, nobody's going to discharge somebody with a less than -- or at one day.

CO-CHAIR ROSENZWEIG: So if a person is -- has a subsequent MI in this interval period, that's considered part of the -- subsequent to the original MI or does the

clock start ticking again?

CO-CHAIR CURTIS: Right. So there's only one index admission per calendar year because it's a 365-day follow-up. So you can't have more than on index admission. The subsequent MI would either follow up within the 30-day measure or it would be counted as part of the outcomes for the 31 to 365.

I think there will probably be different criteria for the acute care, to address that specific issue. But practically one admission with an MI per patient per year.

CO-CHAIR ROSENZWEIG: So the index admission could be within the 365-day period of a previous MI?

CO-CHAIR CURTIS: It wouldn't count as an index in that case.

CO-CHAIR ROSENZWEIG: It wouldn't count?

CO-CHAIR CURTIS: Would not. And so in this case -- well, let me think about that. I believe it's -- and you'll have the

measure developer comment on that, but my understanding was that it was within a calendar year, one index per patient.

DR. WEISS: That is correct. It is only a single event during a calendar year period. So in the example that's been talked about, the second even would group with the very first event and would not count as a new AMI index event for this measure.

CO-CHAIR ROSENZWEIG: Suppose a person had an MI in September of 2009 and then had another MI in March of 2010. The March one would be the index for that year, because it's in a new calendar year?

DR. WEISS: I may have missed the point -- what calendar year timeframe are you measuring? If you're looking at 2009, if you're identifying events during calendar year 2009, then the event in September would be your index event. The index -- the event in 2010 would not get counted as a new event for that individual.

CO-CHAIR ROSENZWEIG: Okay. Just wanted to clarify.

DR. MARWICK: I think the question is, if you're -- if the year you're examining is 2010 and somebody had an infarct in 2009, do you still count the 2010 infarct?

CO-CHAIR ROSENZWEIG: That's my question.

DR. MARWICK: Right.

CO-CHAIR CURTIS: So I would say for the 2009 measure, the first one would count. And in the 2010 measure, the second admission would count as a new index?

DR. MARWICK: The second admission would count -- so it would count but, in fact, the patient would have had a previous MI, correct?

CO-CHAIR CURTIS: Correct. The only MI's that are excluded are those within the 30 days and immediately preceding that admission.

DR. WEINTRAUB: Maybe our

statistical consultant can comment on that. I mean, it's not going to happen all that often, but by doing this, you have a problem with the interclass correlation, and at least it should be counted for.

CO-CHAIR CURTIS: It probably isn't terribly important.

DR. WEINTRAUB: How statistically can you handle people that are showing up in a measure twice with an index event or three times?

CO-CHAIR CURTIS: But let me be clear. They're not counting -- assuming that the measure is at the calendar year, they're not showing up in the same measure in the same calendar year. And that's true of all measures that we have for outcomes, process, et cetera.

MR. ALZOLA: So do we just not worry about it at all?

DR. WEINTRAUB: I don't think it would happen in a proportion high enough that

it would make any difference.

MS. CLARK: Also I was thinking about that as well, in terms of, you know, they're looking at various -- well one stratification, I guess, people with congestive heart failure. Well, they may have more higher costs. You know, the -- I guess the HCC score is probably adjusting for previous MI, I'm assuming. So if a patient had another previous MI, in another year you may think that their costs would be higher if they have one in the year you're measuring, I guess. So maybe it's accounted for the in HCC adjustment method? I'm not sure.

CO-CHAIR CURTIS: I would think, to a certain extent, yes.

DR. WEINTRAUB: Okay.

CO-CHAIR CURTIS: So then there's additional step three, identifying patients with other exclusion criteria. And in this case, I think most of them are reasonable and take their lead from other measures, excluding

patients with active cancer, end stage renal disease, liver disease or HIV AIDS conditions which would be likely to be associated with increased costs and maybe not being -- complicating the -- or making a more heterogeneous population.

So that seemed reasonable to me. But the one that I did focus on was discharge to a -- excluding patients discharged to a skilled nursing facility at the termination of the index hospitalization. And the rationale that's provided is that, I think, difficulty identifying and characterizing the costs associated with this population, and -- I'm trying to find the specific verbiage on it. There was a rationale. But to me, that seemed like a very dicey proposition, to systematically exclude a fairly significant population who would be expected to use a lot of resources and may have unintended consequences in the worst-case-scenario of preferentially sending people out to avoid

being measured.

MS. CLARK: I think maybe what the issue is with that, though, is that once you get into skilled nursing care, for Medicare, for example, they don't cover very many days in skilled nursing. And then it will transfer over to Medicaid. So you're looking at a different claims database. You may be -- if Medicare were to implement this, for example, they're only going to get their costs and not those that get transferred into a Medicaid program that would be paying for the skilled nursing care.

So I don't know if that's why or not. That wasn't specifically laid out, but --

MR. ALZOLA: May I? I think the issue with excluding patients due to mortality or transfer to another facility is that it's initial censoring. You don't know the cost after they are discharged from the hospital. Even though that should be attributed to them,

it's not -- it's not known.

And in the case of mortality, just for the fact that they die, the cost is censored that way. So it's --

CO-CHAIR CURTIS: I don't think they're censoring for death at all. But I guess, for me, for this particular measure, for this interval, 31 to 365, you know, there are a lot of patients who would be potentially discharged to short-term rehab which I think would still qualify as a snip. They may be out of that rehab facility at 31 days. So again, I just don't quite understand it. It doesn't feel terribly comfortable.

Can I ask the measure developer to provide the rationale for that?

MR. WEINSTEIN: Sure. The rationale is reflective of the discussion that we -- when they're censored in our acute period, from day 1 to 30, we are fearful that we can't measure resources that are being consumed that are in the skilled nursing

facility. And because this is intended to be a parallel measure, we didn't want to put those individuals back into this post-acute period.

CO-CHAIR CURTIS: Which I guess is just beyond me, as to whether or not that's a reasonable decision.

DR. LEE: An addition note from developer, and that is, empirically it leads to a very small number of exclusions on this particular exclusion. We wanted to err on not giving false information on this one because of the unknown ability to capture the data.

I add one more piece in and that is that this measure is attributable to the -- most attributable to the individual physicians. And so it would complicate it even more so.

So erring on the fact that there would be a small population that would potentially be excluded from information here, particularly in this commercial population, we

just -- it was a conservative exclusion.

CO-CHAIR CURTIS: So --

DR. WEINTRAUB: Well, the problem that comes up, of course, is if there's variability in this. And do we lose our ability to examine that variability why this exclusion? And how big a problem is that likely to be?

CO-CHAIR CURTIS: So I guess based on the fact that it's a small population in this commercial data set, that it's probably not going to be that impactful. But I would provide the feedback to the provider -- to the developer that it's something we would want to see more data on as it becomes available. Or perhaps some sensitivity analyses to understand what its potential effect could be, depending on different proportion of patients being discharged to rehab or to SNF.

So moving on, you know, specifying on page 13, this is not an all-resource use measure, this is a -- the outcome is specified

to services that are likely, in the opinion of the developer, related to the care of the AMI patient. And this is really the key, and I think it's probably consistent across all the different ABMS-REF measures. Which I'm just going to use ABMS, understanding that it's not accurate, because it's easier to say.

So if you look at page 13, they give you the DRG-ICD-9 codes, et cetera, that are used to specify resource use in this population in this timeframe. And by and large -- and to develop this, I think they went back and forth with their working group trying to identify the codes that were most likely to be attributable, and that I think it was, again, an iterative process where they added or subtracted codes.

If you look at the highlights, they're capturing all codes with the discharged primary diagnosis inpatient of AMI, unstable angina, arrhythmias, pacemaker placements, cardiographs, PCI's, CABG,

coronary artery atherosclerosis and heart failure, which seems fairly comprehensive on the one hand.

On the outpatients, they relaxed the criteria a little bit in that it's a similar range of ICD-9 and ICD-9-DRG's, et cetera. But that associated in either in the primary or secondary position. And that rationale for that is that the ordering of codes in the outpatient setting is less -- I guess less important, is the word that they used, or perhaps less relevant. So -- and I would say, maybe more arbitrary. But that's a key decision in the characterization of this outcome.

MS. CLARK: May I just make a comment here?

CO-CHAIR CURTIS: Sure.

MS. CLARK: So I think these codes need to be updated for the most -- for the current codes that are in use. I noted some discrepancies in the PCI codes for inpatient

ICD-9 codes. Those are outdated codes and they need to be changed to make them more current.

And also it may be just a cut-and-paste issue, but under outpatient events, those are all inpatient codes. So I mean, those need to be -- at least the group there that we're looking at on the screen, it has all DRG information, which is not relevant for outpatient.

And then I was also curious about maybe adding some additional codes that would be relevant or could be relevant for this population, which would be the use of IVUS and fractional flow reserve, as well as coronary CT angiography. Whether they might want to include those.

CO-CHAIR CURTIS: Right. And those overlap exactly with what I was thinking, in terms that there are outdated codes. I pick up on the CPT -- this is, I assume, hospital outpatient services would use

CPT codes potentially, as opposed to the ICD-9 procedure codes.

And that's actually something that would need to be rectified before we go, because this could not -- this measure could not be implemented using the codes that they've specified.

Developer, can you respond to that?

(No response.)

CO-CHAIR CURTIS: I'm sorry, could you hear us?

DR. WEISS: Yes, I'm sorry, I was on mute.

That would be easily rectified if there was an interest in this --

CO-CHAIR CURTIS: We can't hear you.

MS. TURBYVILLE: Kevin, you're fading out.

DR. WEISS: I apologize. I'm in my last stages of getting ready to get in the

airplane.

And is Dr. Lee here still as well?

DR. LEE: Yes, Kevin, I'm still here.

MR. WEINSTEIN: Okay. So what we're saying is there would be no problem for us to -- I think the word used was "rectify," look at those and get those identified. The -- I'll just leave it there. That seems very straightforward and well-appreciated.

DR. WEINTRAUB: So I want to just comment on sort of the general philosophy in costing. Do you want to cost the -- everything that occurs during an episode, or do you want to try and be specific? There's no perfect answer to that, but I think you see in this discussion some of the problems have been that ensue, if you try and -- if you don't cost everything.

Obviously the problem with costing everything is, you add a fair amount of noise. How relevant is the knee replacement with --

that occurs nine months after myocardial infarction? On other hand, you have problems with both error of including things that you didn't mean to include, not including other things that you should. And then you have problems of how do you deal with that attribution?

So what do you do with the pneumonia that occurs two months after a hospitalization for heart failure? Is that relevant or not? Well, yes and no. There's no perfect answer to that.

And you know, it seems to me that relying on the developers of these to make that decision is probably not where that decision ought to be made. That should perhaps be part of the instructions on how to develop these measures in the first place.

CO-CHAIR CURTIS: I think that it's within the realm of their discretion. They can specify it how they want and we can evaluate it for our biases and preferences.

But I think it's a good point. This is not by any means a global list of all the things that can be direct results of AMI care and the index admission and the part beyond.

And so there were decisions and assumptions that were made here, and I think they depended heavily on the input from their working group. And the question I guess for the group here is, is that sufficient? Is this compelling?

Do you want to get into the results? I think we can look at that a little bit. But I want to look at one specific part which is in the slides, the accompanying slides at page, I think, 10 or 11. Sorry, let me see where I started highlighting. Slide number 9 for -- so it's after the application, there's Power Point slides of some sort in PDF format. And then slide 9 shows the top 20 non-AMI related imaging in the post-acute episode.

And I guess my assumption, this

wasn't terribly well labeled. My assumption was that these were codes that would not necessarily have been entered into the resource use measure. But I want to confirm that.

DR. WEISS: That's exactly correct.

CO-CHAIR CURTIS: All right. And so when I looked at this list, it made me a little concerned in that -- I don't know if you can pull it up, Ashlie, but we'll get there eventually. But you look in -- under codes that were not captured routinely, including in the registry you have "Chest pain NOS," which -- with the cost of 86,000 associated with it. And that's, you know, SPECT imaging used for the evaluation of chest pain NOS.

And that is, to me, even though it's not captured by an AMI or arrhythmia or heart failure code, that's the care of the patient most-MI. And in fact, that's -- I'm

surprised it's that low of a frequency. Because when you're filling out the requisition for a stress test, you click whatever is conveniently -- whatever your eye rests on immediately. And so that's a problem for me. And I'll throw that out for other -- but there are lot of other ones like that. CHF, NOS, shortness of breath, precordial chest pain, all associated with and in this case imaging studies that I think are probably in the appropriate framework for inclusion in the measure.

So I know you did an iterative process, I just wonder if it was iterative enough.

DR. WEISS: If I may respond? Just because I'll be having to step off shortly and handing this over to Dr. Lee and Robin Wagner, you're -- the committee has -- you go to this discussion, this is a central one that we, as a developer, worked through.

We heard pretty clearly from the,

principally, physicians. I'll call it our provider community, because we did have other providers on our work groups. But that the total cost did not seem to have strong face validity in terms of their understanding of care of these issues and also in terms of attribution, such that they did not want to go down the development route.

And what that led to us is to actually go through the iterative process which we felt we got to that pretty clear demarcation where that set of experts were able to say, this is going to include 80, 90 percent of what we need. And there will be some missing but that one would look to say that it's not a matter of missing unless it's a question of, do we sense that there's going to be any directionality that we can build in or argue for why, if this is -- or this little residual is missing, it would be important?

And what they captured was the important stuff and grabbed as much of the

important directs costs as possible. We are pretty committed, because of the way that we developed these, to believe that the providers need to have a set of measures that they feel that they understand and that they feel they can take action on with relationship to the condition under study. And that can be matched to the quality metrics.

So that was a fundamental decision we made early on. What we hope you would look at is, on some of these cost sets, because there is no right demarcation. It will be a gray line as to when to include and not include costs, that we avoided things that were -- we did not miss things that are higher frequency that seem to be related. But on the same hand that we also avoided anything that would lead to a directional bias and not wrestle that one to the ground.

Let me just check with Dr. Lee if there's anything that he might want to add to that reflection?

DR. LEE: No, Kevin, I don't have anything to add.

DR. WEISS: I'm not sure if that's helpful to the committee, but at least you'll -- as you go through our measures, you'll understand that very specific reason why we took that approach.

CO-CHAIR CURTIS: And I understand that every measure that we're evaluating has had to make hard decisions as to how they're specifying it. So -- and certainly this is -- yours is clearly demarcated and you have a rationale. But I do still have that concern. And I guess if there were not as big variation in coding practices, both across regions and by physicians, I would have less concern. But I think that's a pretty wide variation, so it's not distributed at random.

DR. MARWICK: Can I just add a point about this? I'm concerned about the non-MI related imaging that actually includes a bunch of things that probably are pertinent

to the MI. The assessment of mitral valve disease, for example.

CO-CHAIR CURTIS: That are --

DR. MARWICK: Right, yeah.

CO-CHAIR CURTIS: Okay. So of note, and we'll come back when we evaluate the criteria.

But continuing on page 14 -- and I realize that we've got 36 pages to get through here, so we're not really going to be on track. But hopefully, it's generalizable to a lot of the measures, and so this is a worthwhile investment. But -- I know, if you feel like I'm hogging the microphone and need to move on, just kick me under the table.

In terms -- okay. So regarding -- so that takes care of the inpatient and outpatient surgeries, procedures, et cetera. Moving to pharmacy services.

They're again, trying to apply restricted -- so they're not looking at all the pharmacy services utilized, but those that

are relevant to AMI care. And so they specify beta blockers, ACE inhibitors, ARB's, Plavix with the lower medications and nitrates. And these are all, I think, reasonable choices. What's more instructive is what's not included and also brings up the issue of the maintenance of these measures, which is going to be substantial accounting for different changes in coding over time as well as changes in the pharmacologic treatment of these patients. So prasugrel is not in here, which probably wouldn't have been in the '06, '07. Ranolazine, I think however, would have been, which is a very expensive treatment for chronic angina.

And then notably is the absence of the diabetes medicines, which I assume is a cognizant choice. But to me, aggressive care of a diabetic patient with an MI is sort of critical in the overall assessment of the care of the patient.

In addition, there's a long list

that you can see -- maybe enlarge a little bit -- of injectable medicines which are broken out. But I wouldn't expect any of these to be applied outside the acute care setting. So I'm not sure how that's different than -- or how that would not be included in the bundled payments to hospitals for inpatient stays.

Any other comments from the group before I ask the provider?

DR. WEINTRAUB: So the only things that will happen is, if you want to compare groups and you don't include certain things, you see very rapidly what will happen. So if you want to compare diabetics to non-diabetics but you don't include diabetes medications, obviously you're going to create a problem in interpretation.

CO-CHAIR CURTIS: But I'm confused. What's your conclusion from that?

CO-CHAIR ROSENZWEIG: Is it appropriate --

DR. WEINTRAUB: I don't know, it's

a problem. You've got to decide -- you know, you've got to decide what things means. You've just got to make decisions on understanding. My choice would be to include the diabetes medications. I think yours is the same, Jeptha. I'm just pointing out what would happen if you don't.

CO-CHAIR CURTIS: And so let me ask the measure developer just again to explain the rationale further.

DR. LEE: So I'm going to -- this is Todd Lee. I'm going to assume that Kevin may have had to leave us. But the rationale for medication was just like the rationale for all of the other services, in that the advice we received from our clinical work group was to focus on services that are direct -- they think most directly related to care of AMI.

The diabetes medications were not included because we're focused on AMI care here. Now granted, there might be a correlation between having an AMI and having

diabetes. But if there was a differential case mix of diabetics across providers and we include diabetes medications as part of an AMI measure, that's going to potentially bias some of our resource use measures. And for that reason, we focused specifically on AMI-related medications.

Now in terms of the sort of injectable list that's there, part of that was when we went through the data, we identified some codes in our HCPCS claims that were not grouping into our episode because they did not have the relevant ICD-9 code, and the work group wanted to include those medications as part of -- as part of this episode.

So in most cases, I believe those were bundled into a hospital claim, but there were certain circumstances where they did show up in the data.

CO-CHAIR CURTIS: Thank you. So I -- just to push you a little bit on that, though, you know, it seems odd that you're

including lipid-lower medications, which is a Class I indication for that class of medicines post-AMI, but so is, I believe, the -- you know, in terms of the guidelines, they recommend aggressive care with a goal hemoglobin A1c.

So I don't know, I don't know if you can draw a bright line. I think I'd believe it more if you said, well, this is being addressed in the diabetes measure, in sort of a complimentary measure as opposed to excluding it wholeheartedly. I would think that the risk adjustment, if it is robust, would account for the differences in the case mix where the diabetes should be identified most often upstream in that 12 months prior.

DR. WEINTRAUB: It's certainly true that the diabetic subgroup will have higher costs, by numerous studies. There's also, you know, recent data suggesting that very intensive diabetes control might be associated with worse cardiovascular outcomes,

including worse mortality.

DR. WEISS: This is Kevin Weiss again. Can you still hear me?

CO-CHAIR CURTIS: Yes.

DR. WEISS: So I'm able to track this a little bit. But it -- so first there is a diabetes measure that you'll be seeing that actually does capture these costs. And recognizing the impact of blood pressure and lipid management as part of diabetes control.

But keep in mind on this measure that we do adjust for case mix of diabetes, so that a physician who is being evaluated, or at least a -- the output of this resource use measure, at whatever aggregate level that it's used at, will actually be able to balance the fact that, if there are higher or lower patient populations with diabetes, without necessarily having to bring in the diabetes costs, per se, into the specific costs associated with the care of these patients who are post-MI.

So there's a very strong internal consistency in how the working groups wanted this developed.

DR. WEINTRAUB: So let me push on that a little further. Let's say that you want to compare physicians on post-MI care in diabetics. Would that -- and there, the use of diabetic medications may be very important. Then you would say that you couldn't use this measure, you wouldn't use it and you would have to go to the diabetes care measure to do that?

DR. WEISS: No, no. I wasn't suggesting that. I was suggesting that if one was looking at diabetes care, then that was identified in a separate measure activity that you'll be reviewing.

DR. WEINTRAUB: All right.

DR. WEISS: What I'm saying here is that it was viewed that the clinical work group was very -- they were very cognizant, as you would expect them to be of the fact that

a patient with diabetes have higher prevalence of comorbid cardiovascular illness and its principal major outcome.

However, in terms of managing the cost of cases with CAD, that they wanted it to be very specific through those types of costs that were associated with CAD care, and recognizing that the diabetes mix within populations would be able to be managed by an adjustment and risk adjustment model.

DR. WEINTRAUB: But I think there's actually a problem there, in interpretation that's difficult. So that if you have variation in let's say comparing health care systems, not the level of physician but health care systems. And one health care system post-MI really emphasizes good diabetes care and one doesn't.

So at least in that, something is -- we can't adequately look at post-MI care in the subgroup of diabetics if you don't account for their variation in care.

Now you could say that's true of anything, but when it's not really terribly interesting, when you look at osteoarthritis. But the diabetes -- because that's not a usual subgroup. But the diabetes/non-diabetes subgroups are -- that's always of particular interest in considering patients with ischemic heart disease.

CO-CHAIR CURTIS: So I think in the interest of moving forward, we should move this to the parking lot. But you've heard that there is some concern about that exclusion criteria specifically. And I guess I would ask -- because this gets into the next criteria -- is why you didn't stratify the measure by that. You opted to stratify by the presence of heart failure in the 12 months prior, which I think is appropriate because that's a higher risk population, that the risk model may not adjust for completely. But how did you decide just to stratify based on that? Why not other things like cardiac arrest or,

you know, COPD or diabetes?

DR. WEISS: So the -- within the 12-month cycle again of this measure that's a little less than -- we tested on an 11-month cycle, we looked for those --

CO-CHAIR CURTIS: Kevin, you faded completely. I don't know if they've closed the doors on the plane yet.

DR. WEISS: I apologize. Is this a little bit better?

CO-CHAIR CURTIS: Yes.

DR. WEISS: For the -- look within the 12-month cycle of the measure, and what interventions would actually relate to research use and associated outcomes in that 12-month cycle? And the one that was very clear was the comorbidity of heart failure, which was a proxy for the severity of myocardial injury. And there's a strong literature that supports that as being a highly predictive of different outcomes.

That was the only reflection of

severity that we were able to gain from that discussion. It was recognized that diabetes does affect outcomes, but it's really not short-term, it's really intermediate, long-term outcome in terms of its impact. And again, so it wasn't viewed as a need for stratification but rather as a -- built into the risk adjustment, so that it was accounted for but not highlighted.

CO-CHAIR CURTIS: Fair enough. So then I think we're going to start breaking off bigger chunks as we go along here.

The one important thing regarding 8.6 concurrency of clinical events, all the, I believe, ABMS measures are specified as stand-alone measures. They cannot be rolled up into any sort of a composite measure, which I think is important for considering all of these. And seems reasonable if it's, you know, trying to drill down on a particular population, but not exclude the possibility of overlap or conflation across conditions.

Then in terms of -- starting to move to 9, you get into the construction logic. And I'm not going to go into the details of this, but basically they identify the relevant population, identify the relevant outcomes using the codes that they've prespecified, and sort of count them up and apply a standard costing to them, which begins on -- sorry -- I guess 9.7 is where you start to get into it, where they apply to specific types of services of inpatient, outpatient, pharmacy and ancillary.

The costing, they've used a standardized cost approach, which uses either information from DRG's or supplemented DRG's plus flags for major surgery. They use similar -- sorry, for outpatient services, they use an average across the whole population. And I'm a little at odds as to how much detail to go into at this level.

To me, the standardized costs that they were calculating all seemed very

reasonable. And I don't know if other people had concerns based on their reviews of similar measures by this developer.

MS. CLARK: Yes. I just had some questions about exactly how this -- these standardized costs were being calculated. Because they're not -- it's not clear how it's done. So it's really not transparent for me, in terms of if I were going to go replicate this, how would you actually do it.

Especially on the outpatient costs, I'm wondering what they did?

CO-CHAIR CURTIS: So would the developer comment on that?

DR. WEISS: Sure. So we -- for each type of procedure -- so if you took a specific CPT code and ICD-9 code combination, we calculated the average paid amount for that specific combination across our data set, and used that as the average cost for the type of claim that was submitted.

CO-CHAIR CURTIS: And so for that

example, did you derive that average cost in all populations in the measure or just the -- I'm sorry, just the population in the measure or in the entire population in the database?

DR. WEISS: It was across the full data set. Not just -- it was not just limited to the population within the measure.

DR. WEINTRAUB: So basically your standardized costs are average payments?

DR. WEISS: Yes.

DR. WEINTRAUB: All right. So with respect of -- on costs, long ago I remember a lecture hearing, there's charges, payments an costs and they have nothing to do with each other. So one has to watch out. You know, to do something like this, you need standardized costs, but there's no such thing. And any time you're trying to do a cost analysis, you're trying to come up with some kind of proxy for societal costs, which is what you really want. But there is no one measure of that.

And I think in terms of developing measure like this, this is just critically important, you know, for those of us who write and need this kind of literature, you read it and even when you're reading the literature published in the best journals, you always have this level of skepticism. Do I really believe these costs are right? If you compare, for instance, the papers I've written to the papers that Mark Hlatky's written on cost of revascularization, you'll find that his are considerably higher than mine. He's using a different costing approach. Is he wrong and I'm right? Or am I wrong and he's right? No. You know, there's just no perfect answer to this.

If you could say that -- if you could always get the scaling right, it doesn't matter what the real cost is because it's a matter of scaling the different items of resource use. But then at the end of the day, do you believe it? Do you believe these

scales get it right?

CO-CHAIR CURTIS: So I think for purposes of this evaluation, though, I mean, I think the concerns are -- but they made their assumption, they made their decision, and we're just evaluating that decision. But I want to make sure that --

DR. WEINTRAUB: It's not the decision of anybody else.

CO-CHAIR CURTIS: Right. And does anyone else have any specific questions about the methodology they used to get the standardized costs?

DR. HWONG: Can I further understand how the NCQA relative resource use, you know, standardized daily price tables actually factor into this? I'm just trying to understand how the ABMS measure is utilizing those.

DR. WEISS: I'm sorry, I was on mute. The NCQA -- we started with trying to use the NCQA price tables across all of our

measures so that we'd have a single standardized price, but we found that there was a lot of services that happened within our data set that did not show up in that standardized price table. And what we ended up doing was creating our own standardized price sets across our measures. But we still used the NCQA price table methodology for all of our inpatient events.

DR. HWONG: I see. So it's limited to the inpatient facility events?

DR. WEISS: That's right.

DR. HWONG: Okay, thank you.

MS. CLARK: I just wanted to -- I know we kind of -- we were talking about the different stratifications previously, and I think we might have skipped over a little bit of the detail on that. I know that they're just looking at congestive heart failure, and they were saying that the reason that that was chosen was a measure -- as a measure of severity. But I think stratification, while

it was also described as looking at it based on demographic criteria, age, sex, race and there definitely was some literature that supported that there was variation in those groups. And I guess we didn't see anything addressed on reporting in those categories. So I'm just -- or stratifications. So I'm just curious why those weren't addressed.

DR. WEISS: So the primary stratification that our work group was interested in was the work group that Kevin mentioned previously. We are limited in the ability to look across age groups primarily because of the commercial data set that we used to test this out. It might be very important to compare, you know, a Medicare aged population to a commercially insured age population.

And so I think part of the reason behind not going down the route of age comparisons was the relatively homogeneous age group that we have in our commercially-insured

test data set.

MS. CLARK: Well, what about some of the others? Sex and race that seem to fall --

DR. WEISS: So race is very difficult-slash-impossible to identify in the data that we used. And the work included sex as a variable in our risk adjustment model rather than going through a stratification process.

MS. CLARK: Okay. And just one other comment on this. I know you were making mention of the differences in resource use, I guess, or severity levels based on trying to identify STEMI versus non-STEMI, and that's not possible in this data. That's definitely something that needs to be considered then when ICD10 goes into effect because you will have that distinction at that point. So that measure will have to be revised, I would assume.

CO-CHAIR CURTIS: Well see, you

have that in the current data, it's just nobody believes it. And so this stands in contrast, actually, to the Ingenix measure for AMI where they actually do attempt to stratify -- or, yes, I think stratify by severity, and one of the criteria is submyocardial versus AMI.

So again, key decision across the different developers having different decisions.

DR. MARWICK: Could I just ask about stratification based on the initial treatment? As I understand it, the only stratification here is full heart failure. But you might consider the nature of the original intervention at the time of presentation in those 30 days. That would -- presumably you're capturing that in that analysis. But that would port over to this as being an important potential stratification.

DR. WEISS: Yes, we certainly do -- did look at the initial intervention as part

of our acute measure, and as you would expect, found some major differences in resource use across what if somebody had a CABG or a PCI. And our work groups are -- you know, sort of that was their underlying hypothesis. And so that variability was a key component of -- capturing that variability was a key component of our acute resource use.

Now we did not explore those interventions as stratification during day 31 to 365. The work group did not choose to investigate that as a potential stratifying criteria in -- during this time period for AMI measure. And you know, I can't comment on the clinical rationale behind that. They just did not choose to go in that direction.

CO-CHAIR CURTIS: Do you think it might come back to how eventually it will be harmonized with the quality measures, in that for the quality aspect you're not going to want to adjust for things that could represent complications of care during the index

admission. The index event here is the admission for MI 30 days previously. So I think they probably would like to avoid adjusting for anything that occurred in that interval from admission to 30 days that could be a complication.

MR. ALZOLA: I have a question. If your study finding by heart failure, how can you have a coefficient in your risk model for heart failure?

DR. WEISS: Yes, I mean that's a good question. That was part of our risk adjustment calculations over the whole population. When we reported out by stratification of heart failure, we reported it out after our implementation of our risk adjustment model, simply reporting heart failure versus no heart failure patients. The risk adjusted calculations were done on the population as a whole.

MR. ALZOLA: So your risk adjustment, your study find on the report and

not in the model?

DR. WEISS: That's correct.

MR. ALZOLA: Okay.

CO-CHAIR CURTIS: So moving on to the next segment which is attribution. The -- let's start on 28. There's a fairly clear plan for attribution to the physician level which says, dependent on thresholds for the proportion of patients -- proportion of encounters provided by a single or multiple provider.

So if you look at S.11.1, if a single provider is providing at least 70 percent of the episode's E&M's during that time frame, is there E&M's for AMI-related care that would be attributed to that single physician? If however no one meets that threshold of 70 percent, you could have multiple attribution across different physicians if they were all physicians who had 30 percent. So up to three physicians could have that patient's costs attributed to them.

And then if nobody has at least 30 percent of the E&M codes, then it is not attributed to anyone.

And so I recognize that there are assumptions and -- made in this, but again it seemed fairly reasonable, as good as any other method of attribution that I'd seen.

CO-CHAIR ROSENZWEIG: Could I ask a question about that? So I mean, the number of E&M codes may still represent a very small proportion of the total costs. I mean, if the patient is seeing -- if the patient is seeing someone as an outpatient provider, they could rack up a lot of E&M costs, but they might represent five, ten percent of the total cost. If there's another person who's doing the -- this CABG during that time, or if there's another person who's doing the radiological procedure, you know, the whatever the -- whatever -- you know, spiral CT scan for whatever it is.

So the question, is that a fair

attribution? And then when you have -- if you have less than 70 percent -- if you have a bunch of providers who have less than 70 percent of the total, are you attributing the total costs to each of those people or are you -- how do you split it up?

CO-CHAIR CURTIS: My understanding is they attribute the whole cost to each of those, as opposed to trying to proportion it out.

But getting to the first part of your question, I think that at the end of the day they're trying to identify someone who is more or less responsible for the AMI care of this patient. And so it's not going to be the person interpreting the spec study, it's not going to be the person doing the cath necessarily. It's going to be the person who's seeing them and making those management or -- decisions.

DR. MARWICK: I think there's a problem there. That is that there's somebody

hiding behind the curtain who's actually -- who is actually charging, potentially charging a lot of money but that isn't showing up in the E&M measures at all. If for example you take somebody who's being managed in primary care but is sent as a consultation to see a cardiologist, and then ends up having a CABG, the decision that drives the cost there is made by somebody else completely. And it's -- and so the primary practitioner is the person who's carrying the responsibility.

CO-CHAIR CURTIS: I don't know, but to me that's sort of what the role of the primary care physician and/or the cardiologist who did the initial consultation is, is in fact to do that. And it should be attributable to them.

Now they do specify -- and I may butcher this, so if the measure developer wants to comment, please jump in. But you know, it can be attributed to across peer groups. So there can be a primary care

physician who's -- to whom these costs are attributed as well as a cardiologist to whom these costs are attributed. So the fact that it's a single measure, a stand-alone measure I think comes into play here, where it makes that feasible. But there aren't -- I mean, again, this is a decision that has to be made.

I guess the converse, if you just attribute it to the actual physician who provided the care, then the nuclear cardiologist or the cardiologist interpreting nuclear studies is going to rack up immense costs, just because that's the part that they're reading. So I'm not sure what the alternative is.

DR. MARWICK: So I think the solution is that, at the moment, of the level of granularity that we have with the data at the moment, this is something that should be attributed to on a group basis, or on a facility basis, rather than on an individual basis.

DR. WEINTRAUB: Then you have problems, of course, follow-up care may not be -- may be spread across the facilities. Extraordinarily difficult. But Jeptha, I think you let them off too easily if primary care physician sees a patient six times and is providing excellent detailed care measurable of this fact that the patient has recurring chest pain. Sent to a cardiologist who caps the patient, sends the patient to a surgeon.

I think that not only are -- you don't get the attribution right, but the problem of the distribution of costs is going to be extraordinarily weighted towards those high-profile events that not only do you have attribution role, but you have problems of the distribution when you're looking at relatively rare, relatively high costs.

I think it becomes impossible at the level of the individual physician, and perhaps doable at the level of the facility. But works best in closed-in systems where all

of the resource use are related to that facility. That doesn't happen most of the time.

DR. HWONG: Yes, hi, I just want to echo that opinion. So maybe the measure developer could think about in future iterations, you know, something where there's more of this sort of shared accountability. I think probably what can happen is, if you do have this primary care physician that's really acting, you know, as his true primary care physician, they could very easily rack up those 70 percent E&M's. And then the cardiologist involved, you know, throughout, you know, wouldn't actually be identified or, you know, have this sort of opportunity for feedback or input.

I understand that they have this, you know, second tier, if it's less than 70 percent and you get 30 percent or something, maybe you'll include -- maybe that will probably grab some more specialists. But

maybe something, again, the measure developer could do is say, okay, you know, there are some attribution logic choices, you know, depending on the philosophy of how you want to implement this.

If you're trying to do things in terms of quality improvement for groups and care coordination, et cetera, it might be good to highlight -- you know, figure out who those individuals are. Because again, if you -- the preponderance is, you know, 70 percent, it's really just, you know, your internist, your family practitioner, et cetera. There may be a specials group that, you know, would not be able to benefit from this information.

CO-CHAIR ROSENZWEIG: Yes, in addition, you know, the patient may be seeing a primary care provider for a whole lot of other reasons, you know, upper respiratory infection, the cardiovascular disease will still be listed as one of the codes for the visit.

And so they may be seeing them for a whole lot of different things. And once they send them to a cardiologist, even if they say -- if they're considered the gatekeeper in a managed care plan, it's really the cardiologist's decision often that rules whether or not these various tests are being done. So I question the issue of, you know -- I mean, and they're in large number. And you -- this particular system may also be attributed to point of service type plans as well, where the primary care physician has no control over whether or not these tests are being done.

CO-CHAIR CURTIS: So just of note, though, in the actual attributions scheme now that they have in the accompanying slides, the majority of these patients were -- episodes were attributed to a cardiologist. So I think it's a fair point that you could, you know, say, well, let's just define all cardiologist care and let's define all primary care, and

attribute one within each peer group.

So there are other options that they could explore. But it seems to work to a certain extent. It has some face validity that the majority of episodes are attributed to the cardiologist.

DR. WEINTRAUB: But that's a problem. You know, most of the time it's going to work out, but some of the time it won't. And do we have a sense of how often it's not going to work out and how often it's going to be nonsense?

CO-CHAIR CURTIS: Well, I think one of the things that's concerning in that same episode is, if you get into the -- and I think it's in the reliability and validity testing -- that in this data, at least 35, 40 percent of cases were not attributable to a particular physician at all. That there was incomplete information about the physician. And so that is concerning.

And I don't know if the measure

developer could comment as to whether or not that was specific to the data set tested or if that represents a global problem that would really be a barrier to implementation of the measure at all?

DR. WEISS: Yes, I can't speak to how the ability to identify providers across multiple data sets. It certainly was an issue in our attribution methodology for testing within this commercial data set. That we were not able to identify an attributable provider with certain claims for a large portion of the claims that we had.

CO-CHAIR CURTIS: So I think that would be something that would need -- you know, since this is attributed to the physician level, that's pretty critical if 30 percent of the claims are not attributable at all. You know, that's introducing much more noise than anything else we've discussed so far.

MS. CLARK: All right. There's

this question, but I wonder if it would make sense to try to attribute it to the physician who actually sees the patient on their initial admission for the AMI and manages their care from that point? I don't know if that's a reasonable thing or not.

CO-CHAIR CURTIS: I think it's another choice. I'm not sure if it's a better choice. Just because there are so many hospitalists who it could be attributed to.

We're slowly working our way through here. We're almost at the end of the 2.A.1 criteria. So only fourteen and a half more measures to go -- thirteen and a half.

(Laughter.)

CO-CHAIR CURTIS: The -- so moving on from attribution and the peer group methodology. I'm going to -- I think we've touched on that enough for this discussion.

They then move into 11.5, 11.6, which is the detail measure, outliers and thresholds which I think is key for all of

these resource use, how are they accounting for very high outliers. In this case, they propose ones arising at the 99 percentile such that any value higher than 99 percentile is set to 99 percentile, and it's a subtly different approach across different developers. But I think at least they've defined how they would approach that. And it seemed, again, I don't think there's a gold standard for saying one Winsorization threshold is better than another, you know.

And then in terms of sample size requirements, they do not specify any minimum sample size necessary for public reporting, which is -- I think gives them flexibility in terms of it, but I guess cause for caution on my side as to, you know, is one or two cases, at the physician level, meaningful in terms of even providing that as feedback to the physician. Does it really impact them or mean anything.

But I would almost give them the

out to say that that just gives them more feedback or more leeway in terms of how they are applying the measure when it actually gets implemented.

DR. WEINTRAUB: Yes, the issue of sample size here is important, and really complicated. And it's complicated because the distributions of costs are going to be so skewed with relatively small percentage of the population having very high costs.

I think this is really extraordinarily difficult. I don't have an answer to it, but I'm worried about it, and I wonder, as a statistical consultant, I'm sure it's something that you've thought about?

MR. ALZOLA: Yes, I don't have an answer, either. I mean, the problem is that, to make the measure useful, you're really going to have to have a relatively large sample size to really estimate the costs. And for many facilities, especially small facilities, they don't see many AMI patients

in any given month.

CO-CHAIR CURTIS: Much less providers. I mean, it's hard enough doing it, you know, half of hospitals admit less than 25 AMI's a year. How many physicians are going to fall into that sum? So a single calendar year, more likely than not, isn't going to -- especially if you're using commercial database. You know, out of 25,000,000 covered patients, only, what, 20,000 MI's were found.

And at the end, once you got down to the attributable level, it was 3,800 or so patients who were included. And so you're getting to very small numbers very quickly.

CO-CHAIR ROSENZWEIG: To what extent can this measure be used for external --

CO-CHAIR CURTIS: Do you want to repeat that?

CO-CHAIR ROSENZWEIG: To what extent can this measure be used for external accountability? I mean, if you can't get the

statistically significant differences between physicians, are there any provisions that this measure would not be used for external accountability, at least on the provider level?

CO-CHAIR CURTIS: I think it's a concern, if you don't have enough cases, how can you be held accountable with statistical power --

CO-CHAIR ROSENZWEIG: Exactly.

CO-CHAIR CURTIS: But let me throw that out to the measure developer. How would you approach this issue of the small number of cases, as well as the overall noise among all the things that we've discussed?

DR. WEISS: I think these are very valid concerns, and we don't have enough information yet to be able to provide a good estimate of the sample size that will be required. I think that will be important for continued maintenance of these measures and understanding exactly how these measures are

being used.

Our initial efforts were focused on identification of the resource use, hoping to get this to the provider level. We acknowledge that there needs to be some additional work around identification of the sample size that's sort of sufficient to be able to provide very sort of robust estimates of relative resource use.

That being said, it may provide some initial benchmarking through, you know, a handful of cases just to give providers a sense of where they lie. You know, the intent is that these would be used as information tools to help to identify variability and resource use. And as such, I think that even with small numbers of cases, there is the potential value for using these measures to identify cases of incredibly high costs or high resource use or what might be driving those, and the variability within there.

DR. WEINTRAUB: So the extreme of

this is pointed out to me by Jeptha, and he's putting it out to me on the AMI post-acute care. And I think it's going to run through all these. It ran through the three I looked at. If you look at inpatient facility costs in particular, you see that the 75th percentile is at zero, 95th percentile is almost $26,000. That's really rather extreme.

So what you're going to have is, until you probably get into -- I don't know, but I would imagine until you get into the thousands, what you're going to have is almost impossible to look at the individual physician. But even the facility, a small facility is going to be almost impossible. Certainly well into the hundreds.

Am I making sense?

MR. ALZOLA: Yes. May I say this, that you perform some simulations --

DR. WEINTRAUB: Yes.

MR. ALZOLA: -- as to how many cases you would need to estimate -- to see the

reality, and wide your confidence in these are. Let's start with 10, 15, 20 and so on, and see what -- if you can get a reasonable sample size with a relatively narrow confidence, you know.

DR. WEINTRAUB: Yes. So I completely agree with that, and that's the kind of simulation that really should be done, really now. That can be done with what's at hand now.

DR. WEISS: Sorry, if I could jump in around this conversation. One thing to keep in mind is that part of what we're reporting are ratios of observed to expected costs. And granted there's a huge degree of variability in the observed costs that we see, that you've pointed out in that distribution, over the average cost of an episode. That is all reflected to a risk adjusted cost, and so now we've got a range of observed to expected ratios.

And granted, there can be a large

range there also, but it can help to reduce the amount of variability we see relative to the huge -- the large skewedness in the cost distribution.

CO-CHAIR CURTIS: So the thing that's defining that, though, is as you note, the thing that's driving costs is, you know, additional revascularization procedures, predominantly on the inpatient basis, right? And that's not randomly distributed, that's distributed by the severity of their disease and the proportion of patients with pre-vessel disease who may not have been taken care of in the first 30 days.

And so it's -- without that granularity of anatomic data, your risk adjustment methodology really can't take that into account. But you've done the best you can with the data you have.

DR. WEISS: Plus -- this is Kevin -- that's the assumption that care is being delivered, let's say appropriately, that there

may be a lot of other revascularization and/or other invasive activities going on that may, in fact, not be consistent with guideline care. And I think that we recognize that's a part of why we're trying to put these measures into practice, to see what that looks like.

DR. WEINTRAUB: Fair enough. I mean, that's what you'd like to do. But the question is, can you pull it off? Because the problem is that you're going to have a variability because of the relatively small number of patients that any one provider, and even most facilities see. And then the relatively small number of revascularizations.

Now if you have an extreme outlier of someone who's doing, I don't know, 20 when they should be doing four, that's one thing. But the problem is that you're going to have this variability just in the stochastics of this. So then how can you get around it? The person who takes care of 20 MI's, most of them will have no revascularizations. The person -

- the people who just have one all of a sudden stand out.

CO-CHAIR CURTIS: And I don't necessarily think that's a problem for this part, right? I mean, it is -- simply, it is measuring -- this is a methodology for defining the costs or resource utilization. It only becomes a problem when you extend it to inferences of quality and value. And that's the concern.

CO-CHAIR ROSENZWEIG: Yes, I would agree with that. And so I -- would it be unreasonable for NQF to specifically require some statement in this context, to make sure that if one of these measures is to be used for external accountability, especially at the provider level, that they need to actually demonstrate that their -- you know, that their statistical basis for it, and rationale for it?

Otherwise, a measure like this could be misused, and in very significant

ways. I mean, especially with all of the paper for performance schemes out there currently.

CO-CHAIR CURTIS: So Sally, I don't know if you want to comment.

MS. TURBYVILLE: I actually deferred to Helen. So the question being, would NQF feel comfortable in the reports or elsewhere saying that, while there's no specifications necessarily on what kinds of statistical properties the measures should meet for public accountability, would we say that users should have a statistical approach, and also be transparent about that when they report these measures, so that there's a little bit more confidence that they have that kind of additional, yet critical --

DR. BURSTIN: I think it's very reasonable for the steering committee to add that.

DR. WEISS: And just to note from the developer, we would appreciate that extra

statement by NQF.

CO-CHAIR ROSENZWEIG: I mean, one of the problems that has come up, though, that we've actually dealt with in the past with diabetes measures, is that even if you specify that physicians should not be held accountable, what happens is that, if the plan is held accountable, they may often on their own put into place paper performance schemes for physicians that are not necessarily based on good data, that can -- in other words, the plans themselves could misuse it.

So I think some sort of directive from NQF with respect to this issue would be extremely helpful.

DR. LYNN: Tom Lynn. I'm from -- I'm obviously not the rule developer, but -- can you hear me now?

But I think this is something obviously relates to all of us. And NQF, through the physician, hospital and quality guidelines, has already made recommendations

about use of cost measures. And that document requires that you do something to show that the decisions you're making, based on cost, are statistically significant. That whatever benchmark you're comparing to, that the measurement you're using shows that you're only making decisions on folks that are statistically different from that benchmark.

And we certainly -- and I'm sure Kevin and his group would join us in saying, we absolutely think they should not be used any other way.

CO-CHAIR CURTIS: I guess the --

DR. PALESTRANT: This is David Palestrant. You know, part of what we've been asked to establish are the reliability and validity, to define that we think that this is, you know, high moderate or low.

But I find it difficult to -- when we have all these different questions, and it's very nuanced data, very difficult to get to. But it really calls for the fact that you

cannot call this reliable unless you know if it's reliable to X or reliable to Y. Do you understand what I'm saying? So it's clearly not reliable if you're judging a physician who's had two patients, but it may reliable for a physician who's had 100 patients. It may not be reliable -- it depends on what you -- you know, it depends on what you're looking at. And this is a very broad -- this is very broad data.

DR. WEINTRAUB: I want to make one more point about statistical significance. It's all very nice, but there's a huge leap from statistical significance to causality. And at the end of the day, we want to believe that the measures that were put forward for everybody in the country to use are not just statistically significant, but when we say something, it's got -- it's never perfect. You know, perfect is the enemy of the good. But some reason to believe that it's causal. That's a very high standard.

DR. WEISS: Just to note, from Kevin, if I may, and that is, I think by the nature of the discussion and the fact that our colleagues from Ingenix are actually exactly as concerned as suggested, this is a more generic issue across the measures. And anything that can be done to address this would be really important to the field.

CO-CHAIR CURTIS: Agreed. So Sally, just let me ask, do you want to pause now, or should we vote on this particular criteria of 2.A.1 or should we try to get through reliability, validity? Which I think we've really discussed on an ad hoc basis, and I think we could get through quickly. But I want to be sensitive.

MS. TURBYVILLE: Yes, I think we should definitely vote on the ratings on 2.A.1. And so we're bumping up against lunch, so it's really up to you guys to decide. It's waiting, but we also want to make sure we give the public a good opportunity to comment. So

our plan was, once you guys close out here, and then want to move to lunch, right before we go to lunch we would open it up to the public comment.

So if you want to try and move through the next section, you know, I think we should go for it if we feel like we're picking up speed here and, you know, don't want to break in between some of the thoughts.

CO-CHAIR CURTIS: Yes, I'd rather just --

MS. TURBYVILLE: Okay.

CO-CHAIR CURTIS: I'm going to do a dictatorship. We're going to keep going until we finish this, so then I can stop talking afterwards.

(Laughter.)

CO-CHAIR CURTIS: -- peacefully go to sleep.

So in terms of -- so then move to 2.A.2, which is reliability testing, demonstrating that the results are repeatable,

producing the same results in a hyper portion of the time when it's based on the same populations and the same time and that the measure score is precise.

So when we get to that, that's scientific acceptability, 1.3 through 1.4, am I right that we're going to go through all these and then vote on all the 2 criteria, or should we vote on 2.A.1 first?

MS. TURBYVILLE: We could vote on 2.A.1 and -- it might be good to vote on 2.A.1 and 2.A.2 together, since they both map to reliability.

CO-CHAIR CURTIS: Okay.

MS. TURBYVILLE: And then also you can take benefit of Carlos at the table, if you want him to provide any overview as well. So I'll leave it to you to ask him as you see fit.

CO-CHAIR CURTIS: Okay. So I think that if you -- it's a little hard to go through, but if you can bring up the slide for

the accessory slide number one that shows the inclusion-exclusion criteria effect.

So they validated or assessed this measure in the commercial available Thomas Reuters data set with 25,000,000 lives. When they're talking about validity testing, they acknowledge that they are primarily focusing or accepting face validity --

The diagram, yes. Sorry.

So I'm sorry, it's slide four.

So when they -- again, this is how precisely specified is the measure when they apply their criteria of continuous coverage, standard NCQA exclusion criteria, age restriction, et cetera? I think that, to me at least, you get from 10,000 patients at the start down to 3,800 patients at the end, whereas there are concerns about the individual exclusion criteria, I think at the end it is precisely defined. So I think that if you replicate this across -- in the same data set or across other data sets, you would

have a similar ability to come to the same cohort.

So from that standpoint, I think it's precisely defined.

And then you get into how they are -- proposed attribution logic or identification of related and unrelated services, et cetera. We've really touched on that I think extensively at this point. Where there are choices that they made in terms of related and unrelated services that they feel represents the majority of AMI attributable care is captured in the measure, with some acceptance of loss based on -- if you keep scrolling down, the unrelated procedures that we looked at before, slide nine. Non-AMI related imaging is an example.

And when you look at the incorporation of the risk adjustment, they don't really provide necessarily data on this, but you know, they are using HUC, which is an accepted risk adjustment methodology specific

to cost, which seemed in only using the 12 months prior for risk adjustment. So that seemed fairly specifically placed.

And assessment of the physician attribution, we touched on this. I think that's slide 14, where applying the 3,700 cases that they had, 47 percent had insufficient provider ID so they couldn't attribute to any physician. Within that 1,500 left over, you had 70 percent attribution to a single provider, 1,100 patients, and then a smaller proportion in which the episodes were attributed to two or three providers, and only half a percent in which there was no provider attributed.

So again, this is -- we could argue about whether or not this is the right form of attribution, but I think it's a precise attribution once they get to a criteria. So if they could fix the whole in terms of identifying physicians, then it could be precisely attributed for the measure --

purposes of the measure.

So then you get into testing results and the findings statement. When they did this, and again, we've touched on this already, but when you start looking at the outputs from the measure, which I think I'm going to call page 17, we get into that issue of how this looks. And so the bottom is the sum of costs across providers. I'm sorry, the sum of costs across patients, the variance. And you see that, indeed, there is significant variance in the total costs assessed across patients ranging from 646 in the lowest fifth percentile to 3,800 or 3,700 in the 95th percentile. But that data is incredibly skewed, based on whether or not the patient had been admitted and/or had undergone procedures to a less or the outpatient facility costs.

So there is variation across, but we've raised the concerns as to whether or not this is, in fact, stable case -- or being

driven by measures that the risk adjustment methodology couldn't account for.

And then getting on to slide 18, 19, et cetera, you sort of see how this would work using region as a proxy for provider, in this case. Among the 3,800 episodes that they were assessing, you can see that northeast, the care is different than it is in the south and west. And you might wonder as to that, because it is sort of the inverse of what we find on population studies. It probably has to do, in my off the cuff opinion, as to that this is the post-acute episodic care. So the west may use earlier care in the first 30 days whereas the northeast may be doing more of the care, their invasive procedures after the first 30 days. But again, that's highly speculative on my part.

But again, the conclusion though is that, in this risk adjusted costing or resource use methodology, there is a variation in the cost as assessed by the ratio of

observed to expecteds.

I'll pause there for a second.

And then they replicated at the state level as opposed to region. And I just want to go down to the bottom, the last slide that they have, which is the sample report. It kind of shows how this potentially could be applied to the physician level. I believe that this is not specified for AMI, it's not using the AMI data, probably I think it might have been diabetes, although that's not specified on this slide.

But using similar methodology, there are differences in observed costs, predicted costs and the observed to expected ratio at the level of the provider, within a certain specialty. And that that can be benchmarked against peer groups, non-peer groups and the national average in a way that you could potentially use this to identify physician level differences and resource utilization.

So again, I'll pause there. I went kind of racing through that, and there were a lot of elements. But again, I do feel like we discussed most of them up front.

Carlos, I don't know if you want to specifically talk about how they used the observed to expecteds in your take on the risk adjustment methodology as a whole?

MR. ALZOLA: Yes, the risk adjustment methodology, it seems they use an appropriate method. They use something of a regression where they could use a log model. Which one of those, you just look at them and see which one works best. It doesn't -- you have to be really practical on how you use that information.

What I didn't see, which I was expecting, was how good the model feeds were. There weren't any -- there wasn't any calibration curve to see where predicting or under-predicting specific reasons. I wouldn't be surprised that, if -- to see that we are

almost under-predicting for the really expensive cases. It's very typical.

And so there were no r squares and not any of that kind of information.

CO-CHAIR CURTIS: Right. In fact, they stated that they calculated the r squares and residual means, et cetera --

MR. ALZOLA: Right.

CO-CHAIR CURTIS: -- but it wasn't present in the application. I don't know if that was how they interpreted the specific criteria out of the application or not. But I think that's something that we would really want to see or need to see by the time of the steering committee review.

MR. ALZOLA: And so there's one more additional comment, with respect to the -- how they calculated the observed expected ratios. They did it on an individual basis, so they calculated the observed for an episode, divided by the expected for that episode and summarized those.

It is more typical to look at the observed for all -- the ratio for all the observeds for a provider and divide by the expected ones. And it's not that what they did is wrong, it's -- it has other properties. But I am not so sure of what the statistical properties of that approach are. But the other standard approach is more -- the statistical properties are well known, and it's really -- it's all -- that information is all here to provide it in that way.

CO-CHAIR CURTIS: So let me ask the measure developer to comment on that particular decision of grouping or calculating individually the observed and expecteds?

You might be on mute?

DR. WEISS: Oh, sorry. I've got to remember to push that button.

Yes, we -- we intended to measure this -- we wanted to assess the variability within individual patients. And so we were interested in observed to expected at the

individual level. But like I mentioned, you know, you can roll this up and the information is available here to calculate it across in any level of measure you'd like to.

We focused on the individual because we wanted to understand variability within individual patients and how that was then attributed to providers. We -- you know, we have all sorts of data that we could have provided on the performance of doing it at the individual level, the performance of our risk adjustment model. So that information is easy to supply to NQF if that's necessary as part of the further evaluation of these episodes.

CO-CHAIR CURTIS: I think that would be good.

MS. TURBYVILLE: Jeptha, could you or the measure developer clarify what's going to be provided so I can put it in my notes?

DR. WEISS: We can provide data or information on the performance of our risk adjustment models that the panel is asking

for.

MS. TURBYVILLE: Great, thank you.

MR. ALZOLA: One thing that would be important to see is what were the candidate variables for the risk model and which are -- and how you ended up selecting the ones you ended up selecting.

DR. WEISS: Okay.

CO-CHAIR CURTIS: So I think we should move to the voting component, starting with 2.A.1. Let me first open it up, does anyone else on the panel have any specific comments or requests for clarification?

(No response.)

CO-CHAIR CURTIS: So for 2.A.1, you can see there, you know, all these different microcriteria. I don't know if -- and I think we -- it will probably just depend on your individual preference and take. You can sort of assess this as sort of a rolled up, like global the average criteria or you could apply -- I think if you're that

concerned about any one of the individual criteria is not met, that that may be a killer for a fatal flaw for you for the measure. And I don't necessarily think we have direction from the steering committee as to how to apply that.

MS. TURBYVILLE: So this subcriteria does apply to the entire specification. So you're right, if there are certain components of it, for example, the clinical logic, that you feel is not met, and then therefore -- and it's met in such a way that the specification is not precise enough, we would want that to be reflected in the criteria. But we will also ask for rationale. So in particular, for low and moderate, I want to take some time so that for adjustments that the developer can make, they can. And otherwise, we're capturing that information.

CO-CHAIR CURTIS: And when you say we were going to provide the rationale, who's going to provide the rationale? We've been

talking for two and a half hours, but are we doing that offline or at real time?

MS. TURBYVILLE: I think we can summarize some of what we've heard, and then we'll pause and ask the TAP to let us know if we have missed anything.

So for example, we heard that the codes need to be updated to the most recent version of CPT-ICD-9. So we'll do that after you vote and rate. And if we miss anything that is, you know, pertinent to this rating, you can provide it to us at that time.

CO-CHAIR CURTIS: So, perfect. So then moving forward on 2.A.1, the measure is well defined and precisely specified so that it can be implemented consistently within and across organizations and allow for comparability.

And go ahead and vote.

So it's quite heterogeneous with one high, four moderate, two low and one insufficient. It's a little hard to say, I

don't know, do we average that or --

MS. TURBYVILLE: No, the steering committee will see the exact frequency in the number, so we don't attempt to try and create some kind of overall.

CO-CHAIR CURTIS: Okay. So we then summarize kind of what we heard?

MS. TURBYVILLE: Yes. So some of the things I heard that need to be updated are the length of the data required needs to be aligned with the measurement time span itself. Some clarification, how the standard prices are approached, a little bit more clarity and transparency around that. There -- sorry --

MR. AMIN: There were specific concerns around -- I'm just going to add with you, if that's okay?

MS. TURBYVILLE: Sure, go ahead. Yeah.

MR. AMIN: Specific concerns around the exclusions -- or on the exclusion criteria of excluding patients to the skilled

nursing facility.

MS. TURBYVILLE: Yes, and there was some question about the Rx, and so response back, that is the pharmacy response back from the measure developer about a little bit more in detail about how they're going to be updated through maintenance, especially the pharmacy codes, but codes in general. As well as the HCPCS that were included, a bit more rationale around that.

And there was some question about the sample size recommendations provided. Some of that from the TAP came back to NQF to think about statistical criteria, or at least statements about encouraging or requesting that users of these measures are providing -- are using sound statistical approaches.

There is some missing information on how the model fits for the selected risk adjustment approach, and so there's a request for that information to be submitted, along with a description of the candidate variables

that were examined and how they were selected in the final risk adjustment model.

MR. AMIN: The only thing else I would add is, there was a strong discussion around the attribution approach on concerns of purely defining attribution, based on E&M codes.

Is there anything else that we should --

CO-CHAIR CURTIS: It seems like a lot.

MS. TURBYVILLE: And so, you know, we'll facilitate this with the measure developers and they'll determine how to respond back to all of you.

Thank you.

MR. AMIN: Yes, there -- I would just add one more. There was discussion around the stratification approach for race and sex. So that was added into the discussion.

CO-CHAIR CURTIS: Okay. So for

criteria 2.A.2, reliability testing, demonstrates that the results are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same period. And that the measure score is precise.

And if we can go ahead and vote on that.

Again, sort of a normal distribution around moderate.

(Laughter.)

CO-CHAIR CURTIS: And again, do we need to summarize the feedback? I don't there was anything specific to this.

MS. TURBYVILLE: Yes, I think we covered it in the first.

CO-CHAIR CURTIS: 2.B.1, the measure specifications are consistent with the evidence presented to support the focus of measurement under criteria 1.B. The measure is specified to capture the most inclusive target population indicated by the evidence,

and exclusions are supported by the evidence.

Most moderate; six moderate and two high.

2.B.2, validity testing demonstrates that the measure data elements are correct and the measure score correctly reflects the cost of care or resources provided, adequately distinguishing higher and lower cost or resource use.

Two high, five moderate and one low.

2.B.3, exclusions are supported by clinical evidence or supported by sufficient frequency of events. Its measure specifications for scoring the included computed exclusions so the effect of the measure is transparent. And I don't think it's applicable here, patient preferences are taken into account.

Go ahead and vote.

Two highs, four moderates and two lows.

For criteria 2.B.4, for outcome measures or other measures, i.e., resource use, when indicated and evidence-based risk adjustment strategy as specified, and based on patient and clinical factors that influence the outcome but not related to disparities in care or qualify of care and are present at the start of care.

Go ahead and vote.

Three highs, three moderates, and two lows.

So for -- we're on 2.B.5, right? Data analysis demonstrates -- data analysis demonstrates methods for scoring analysis of identification of statistically significant and practically/clinically meaningful differences in performance.

Go ahead and vote.

Four moderate and four low, no high.

2.B.6, if multiple data sources are specified, demonstration that they produce

comparable results. I believe it's not applicable. Do we need to officially vote on that?

MS. TURBYVILLE: No.

CO-CHAIR CURTIS: 2.C, if disparities in care, have measure specifications and scoring -- have been identified, sorry, measure specification scoring analyses allow for identification of disparities through stratification.

Go ahead and vote.

Two moderates and six lows.

So moving towards usability which -- I don't know, should we stop?

MS. TURBYVILLE: Yes, I think we should stop.

CO-CHAIR CURTIS: Okay.

MS. TURBYVILLE: Because of the two moderates and six lows for stratification for disparities, we may want to either after we get through this measure, come back and see if this is going to be an issue throughout all

the measures. And as I think it was mentioned before, whether or not it's really applicable to resource use measures. So you know, maybe we can move through the rest of this measure and then go back to this, because my sense is it's going to be a recurring issue.

CO-CHAIR CURTIS: Okay.

MS. TURBYVILLE: All right, so let's open it for public comment quickly, right?

Operator, if you could please open the line for any public input or questions at this time, we would appreciate it.

OPERATOR: Thank you. If you have a comment or a question at this time, please press star-one for an open line. Again, that's star-one if you have a question or a comment at this time and we'll pause for just a moment to give everyone a chance to.

(No response.)

OPERATOR: And just as a reminder, it's star-one if you have a question or a

comment at this time.

(No response.)

OPERATOR: And it appears we have no public questions or comments at this time.

MS. TURBYVILLE: Thank you. We're going to break for lunch now.

CO-CHAIR CURTIS: So it's 12:52 now. Should we reconvene at 1:20? Helen says 1:30.

(Whereupon, the above-entitled matter went off the record at 12:53 p.m., and resumed at 1:22 p.m.)

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

(1:22 p.m.)

CO-CHAIR CURTIS: So I think we're going to get started again. I hope everyone's fed and slightly rested. We're still on the first measure, but we are in the home stretch.

And just to orient people who may be on the phone as well as the measure developers, we're not going to -- we're going to deviate from the current agenda, based on a family emergency from one of the TAP members, and so we're not going to consider the -- what number was that, Sally?

MS. TURBYVILLE: 1558 will not be done this morning.

CO-CHAIR CURTIS: So we'll probably -- we moved it up and now we're moving it back probably to tomorrow at the earliest, if not later. Instead we're going to move up to 1593 which is the ETG-AMI resource use by Ingenix. But first we need to close on the first ABMS measure.

So we were moving along with the voting, and we were at the level of usability when we broke. Which we didn't necessarily review in great detail, but it could be reviewed quickly.

As the measure developers note that there is currently no -- it's not in current use, I believe -- on 38, yes, 38 of the measure specifications. And it is basically the measure developers have said in all these criteria that they are in the process of assessing the usability of the measures and have funding to do that, but have not as yet completed that.

So I think we can -- in the absence -- there's not much to discuss, but if people feel comfortable voting, otherwise if you want to discuss any aspect of it, I don't know what the direction or guidance would be from NQF as to that. I would assume that you would have to vote low, at least that was my interpretation, if there was no evidence to

support it, but I don't --

CO-CHAIR ROSENZWEIG: Or insufficient.

CO-CHAIR CURTIS: Or insufficient? Okay.

MS. TURBYVILLE: Insufficient is a little bit more in line with them not having submitted anything or acknowledging that it hasn't been done yet.

CO-CHAIR CURTIS: Okay. So why don't we then move on usability and the individual criteria are almost irrelevant, since we're going to do insufficient. But measure performance results are reported to the public at large in national community reporting programs by the time of endorsement maintenance review. Exceptions considered if there is evidence that the measure performance results are available for public reporting and that the use of the measure has benefited the public.

So go ahead and vote on that. And

again, insufficient is four, not applicable is five.

And similarly, so eight insufficient.

CO-CHAIR ROSENZWEIG: Everybody's paying attention.

CO-CHAIR CURTIS: And hit the right button.

3.B, performance results are considered meaningful, understandable and useful.

Again, go ahead and vote.

I think we can do the next one in under ten seconds.

3.C -- I'm sorry, so one low and seven insufficient.

For 3.C, data and result details are maintained such that the resource use measure, including the clinical instruction logic, for a defined unit of measurement can be decomposed to facilitate transparency and understanding.

So two moderates, one low, and five insufficient.

And 3.D, the measure specifications are harmonized with other measures or differences in specifications are evaluated to be justified. And I think this should be a not applicable or we shouldn't vote on it.

MS. TURBYVILLE: At this point, it would be not applicable. We recently instructed the measure developers that they did not have to try and harmonize with currently endorsed measures. As we examine other measures in the future within this project, then we may ask them to harmonize. So at this point, this is not applicable.

CO-CHAIR CURTIS: I would just hit -- yes, I would vote. We have 35 seconds.

Okay. So under feasibility, again, I think we have touched on all the criteria that would be used to vote on this. If I can open it up, I think the fact that all

the data elements are encaptured by electronic claims submissions is fairly straightforward.

They do, in section F3 at page 40 of the PDF discuss some of the susceptibility to inaccuracies, errors and unintended consequences, which I do think we've discussed in the previous session. I would say that they didn't really explore unintended consequences in the application, but I think we've identified some instances where that could happen.

DR. WEINTRAUB: You know, another thing is that really you have to worry about in claims cases is misclassification. So I think that would -- I'm sorry.

The other things we should think about in analysis of claims is misclassification. And we really haven't discussed that here, but I think in terms of feasibility, it would probably be a part of it.

CO-CHAIR CURTIS: And then the

final section, 4.D of F.4 is the data collection strategy. And you know, they included some information, sort of like lessons learned in the process, but it's a little hard to gauge this. The specific criteria is data collection and measurement strategy can be implemented as demonstrated by operational use in external reporting programs, or attesting to not identify barriers to operational use.

And I think this -- outside of the scientific acceptability, from purely a feasibility standpoint, at least using the claims data and assuming they fix the ICD-9-CPT codes, I think it is certainly feasible to apply this.

Open that up for discussion?

(No response.)

CO-CHAIR CURTIS: Okay. So why don't we go ahead and vote on the elements of feasibility, 4.A. So for clinical measures, the required data elements are routinely

generated and used during the care delivery. I voted high on that, as it's all electronic claims submission.

Five high, three moderate.

4.B, the required data elements are available in electronic health records or other electronic sources, which I think is fairly straightforward. I voted high on that as well.

Five high, three moderate.

4.C, susceptibility to inaccuracies, errors or unattended consequences related to the measurement or judged to be inconsequential or can be minimized through proper actions, or can be monitored and detected.

And I felt like this was low, based on my interpretation of it.

Go ahead and vote.

One high, two moderate and five low.

Finally, 4.D, data collection and

measurement strategy can be implemented as demonstrated by operational use and external reporting programs, or did not identify barriers.

I voted that as moderate, but that was for lack of a more informed choice.

Go ahead and vote.

So five moderate and three low.

So with that, our voting is completed on the first measure.

(Applause.)

CO-CHAIR CURTIS: Thirteen to go.

(Laughter.)

CO-CHAIR CURTIS: So as I noted, we're going to go to the Ingenix measure of AMI. And so that's 1593, and Dr. Marwick is, I think, going to be the primary lead, and I'm the co-lead on that, or co-reviewer.

DR. MARWICK: So this is a measure that's labeled as being for acute MI, but I, I must say, struggled with it a little bit as to whether these all were going to be acute

infarcts. In particular, some of the description of the physicians looking after the patients included primary care physicians, which seemed inconsistent.

So I guess in the interest of time, we should go through it.

CO-CHAIR CURTIS: Let me interrupt there. I think we're going to start with a presentation by Ingenix, is that correct?

DR. MARWICK: Okay.

MS. TURBYVILLE: Yes. So we would like to allow Ingenix about three to five minutes or so to reintroduce the approaches in this measure, if you like, Tom. It's up to you.

DR. MARWICK: That's fine.

DR. LYNN: Thank you.

I'd just take a few minutes to talk about the rules in general, this rule and ones coming forward.

First I'd like to thank NQF and the technical advisory panel for taking time

to review the measures submitted by Ingenix, and we appreciate your thoughtful evaluation and feedback.

The measures submitted by Ingenix and under review today are based on our episode treatment group methodology. This methodology consumes administrative claims and creates case mix and risk adjusted units of analysis around diseases and conditions. The methodology is table driven, which allows for easy maintenance and change to the clinical content leveraged by the method.

Although the features of episode treatment groups related to the diseases today have been extracted for evaluation, the methodology groups claims to all diseases and conditions.

In most cases, a measurement is accomplished by aggregating actual utilization of episodes in the numerator of the measure and case mix and risk adjusted expected utilization the denominator, whether the

measure of utilization is dollars or something like emergency visits.

And that's really all I have to say.

MS. TURBYVILLE: Great, thank you.

DR. MARWICK: Thank you. So we'll proceed going through the format.

The first section relates to the items about whether the post-infarct population is worth of this kind of measure, and I think the case is made very strongly for that.

The second relates to the performance gap, and there is a performance gap, I think, but what I struggled with was in the documentation. That wasn't put as forcefully as I guess it could have, and there were quite a number of sort of generic statements about that.

I would say parenthetically that I think it's very challenging to do this kind of work from this data set. You know, the kind

of things you might be more interested in would be, for example, door-to-needle time and so on, which obviously aren't available. But possibly could be configured if we think about some of the alternatives that were presented earlier for gathering data.

Just in relation to that as well, there is a statement there about CAD episodes, there's a distinction between subendocardial and Q wave infarction and STEMI, but not further detail. And I think that's something we could discuss a little bit later.

Then in relation to the purpose, I think the purpose is pretty clear. But the resource use category information I thought was somewhat limited here. There were reference, for example, to ambulatory care sensitive conditions, which again was irrelevant to this.

So really, to summarize the first component of this, about the importance, I think the statements that were made about it

being important condition were well made. The rest of the information I think we all know from our background knowledge is relevant. But the defense of this in the actual document was not so good, in my opinion.

So in terms of scoring, I gave this a high for impact, and medium for performance gap and purpose, and for resource use scale category, I felt that that information was limited.

MS. TURBYVILLE: Any questions or input from the other panel members?

CO-CHAIR CURTIS: I just want to say, I did appreciate the fact that they put in the empiric evidence of variation in costs derived from the data in this section. I just thought that was useful to have it on hand.

DR. MARWICK: Yes. So are we going to vote on this?

MS. TURBYVILLE: We can. If there are no more questions or input, we can go ahead and vote on this, these subcriteria.

MS. CLARK: I guess I have a question, because I didn't get a chance to read this one.

But I'm just curious, how does this differ from some of the other AMI measures? I mean, what's the definition of an episode?

DR. MARWICK: Well, that's what I alluded to in the beginning.

MS. CLARK: Oh, okay.

DR. MARWICK: We'll get on to talking about that later.

I really struggled with that, as to whether that was an acute presentation with infarction or whether that was somebody with a history of infarction presenting to outside of a hospital stay. And in particular, I was confused by the involvement of the primary care physicians in that process as well. And it wasn't clear to me. I don't know, maybe the developers could help us with that?

DR. LYNN: I think we can help.

Well, the -- in a lot of these processes, we're trying to avoid looking at utilization and having utilization drive severity or drive certain case. But I think -- and so that's why you see AMI being defined, but not necessarily an admission.

DR. MARWICK: Yes, I think this is probably pertinent to the next component, but we may as well discuss it now because, to me, this was the fundamental problem with this. That if we're talking about acute infarction, then particularly in relation to the costs that are incurred, the defining it at the time of presentation is critical.

If somebody has had, you know, an infarct last year or an infarct while they were being looked after by a previous carrier, and then their first -- your first knowledge of their infarction comes from somebody coding it, you know in a non-acute setting, I think they're two completely different scenarios. And I think that plays out enormously into how

you judge cost.

CO-CHAIR CURTIS: Let me, before we get too into that, I think maybe we should just close on importance, because this is really the heart of this particular measure, is how you're defining the episode. So I would just start -- let's table that for just a few minutes --

DR. MARWICK: Okay.

CO-CHAIR CURTIS: -- and get that one vote out of the way quickly, and then we can spend some more time on that.

So let me just, everyone grab their keychains, unless there's more discussion as to the importance.

So for the 1.A, the measure focuses on national health goal and priority. And Tom, your preliminary was --

DR. MARWICK: My preliminary was a high for that.

CO-CHAIR CURTIS: Go ahead and vote.

So that's eight highs.

For 1.B, the demonstration of resource use for cost problems and opportunity for improvement.

DR. MARWICK: So my preliminary was a moderate for this, based on the information presented rather than background knowledge.

CO-CHAIR CURTIS: Go ahead and vote. Sorry, I don't know if you're waiting for me to say that or not.

So for 1.C, the purpose or objective for the resource use measure and construct for resource using costs are clearly described. I don't know if we've discussed in sufficient detail to vote on that.

DR. MARWICK: So what I struggled with here was the interaction with exactly who we were describing, and that's the reason I thought that discussion was pertinent to this part. You know, if this is about judging the costs of an acute episode, then as configured,

this is not appropriately set up.

If it's for judging people who have had an MI in the past, then it's also ambiguous because I think it would include people presenting acutely. So I really struggled with this, and I gave this a low.

MS. TURBYVILLE: So just to add for context purposes, so in the importance, you can think of it as they describe this, if you recall, later on in the validity section, there's a question about whether or not the specifications match the intent of the measure as they've described it.

DR. LYNN: I don't want us all to torture ourselves and each other over this. I mean, it's very clear to me, this is better as an event -- as considered by the TAP, is evaluated better as an event. And it's not how we did it.

And you know, we can go through the voting, but it sounds like we're on the wrong track here.

CO-CHAIR CURTIS: Well, let's finish up the importance. And then as we go through the scientific acceptability, if we kind of reach an impasse, you know, as the co-developer, I shared the exact same concern about the definition of the episode.

But I think for measure intent, what I think we should go by is really just what's written on IM3, which is, you know, the intent of the measure and its components is to support the understanding of opportunities to improve the efficiency of healthcare, in particular for patients with selected conditions, and reducing unwanted variation.

And then secondarily, as a step towards the estimation of value delivered by individual providers. So I think that's the intent of the measure that we're voting on. And then come back to that later and say, "Did they achieve that based on the specifications of the measure?"

So did you want to revise that to

--

DR. MARWICK: Okay, I'll revise that to a moderate. Thank you.

CO-CHAIR CURTIS: Go ahead and vote on that, unless there's more discussion.

One high, six moderate and one low.

And then for 1.D, the resource use categories that are included in the resource use measure are consistent with and representative of the conceptual construct represented by the measure. Whether or not the resource use measure development begins with a conceptual construct or set of resource service categories, the service categories included must be conceptually coherent and consistent with the purpose.

And I think then I would allow the -- your estimation of the measure. And again, we said up front that we might come back to this down the road. But --

DR. MARWICK: So I gave this a

low.

CO-CHAIR CURTIS: Does anyone have any questions about how to vote? I mean, it's sort of a tricky thing. If you're looking into the future to what you might do. I'm still not sure about the placement of this particular voting category.

DR. HWONG: Is this more of like, you know, the categories of data and the way they classify them, you know, as listed, versus the -- you know, versus like when they dig down into the actual codes, which it sounds like individuals are having, you know, some concern with.

But you know, in terms of the general categories of resource use, we should be voting on, you know, what's sort of listed on a high level, is that right?

MS. TURBYVILLE: That's right. So this isn't about the detail. Again, the entire importance section is not about the specifications as written, it's about the area

focus, the type of categories of resource use that they are proposing to measure. To the extent that it might be informed, again, later one, we can understand that. But it really is have they listed a comprehensive set.

As I said, it's really a check box, but we also allow for others, presumably there could have been opportunity for other resource categories to be presented that just didn't really jive with the intent of the measure. So I'm trying to think of an absurd example, but let's say the intent of the measure was to look at diabetes. And in the other category -- resources for diabetes in other category, they mentioned capturing something completely off the wall, I think they would potentially -- or are missing very key components that I think would change their rating results.

CO-CHAIR CURTIS: So why don't we go ahead and vote on that, then.

That's one high, three moderate

and four low.

And that closes us on -- unless do we need to define what we would say back to the measure developer?

MS. TURBYVILLE: I think it would be helpful, especially with so many lows on the last one. If there is some kind of justification or rationale.

DR. MARWICK: Well, I think it's very difficult to define resource measures unless you hook it into specifically what you're seeking to study. And I find that ambiguous at the moment.

CO-CHAIR CURTIS: Any other feedback?

DR. HWONG: I mean, I guess the only thing from my perspective -- again, I think it's interpreting the question and, you know, how you want to look at it. But when I looked at the categories of, you know, inpatient services, ambulatory, which you want to capture, I thought the broad -- you know,

pharmacy, E&M, procedures, surgery, I think the categories seem comprehensive. Again, we'll probably get more into detail about the definition in there. But I didn't think that there was some actual like place of service or category that was, you know, obviously lacking.

CO-CHAIR CURTIS: I agree. I think we probably don't need additional feedback to the developer because I think it will come in the next section.

So why don't we move on to scientific acceptability?

DR. MARWICK: So this is really the fundamental area, I think, which I struggled with.

So first of all is the definition of exactly what is being studied. And then secondarily, if this is an acute episode, then there are a bunch of things that you would want to have in the description of risk that I think are limited at the moment. In fact,

I think the only risk verification is really between STEMI and non-STEMI.

So I would imagine, for example, that previous infarct, valve disease and so on would be important modulators of risk.

I'm not sure I completely understood the pharmacy benefits, but I would -- there's a mention that the incorporating pharmacy benefits has been avoided. And I think that's potentially problematic.

So my other comments there are really related to lack of sophistication and understanding risk. There are no means to allow an episode to shift to another episode treatment group, which I think might be pertinent for some subgroups, particularly ones that are infrequent and cause major increments of cost. For example, the post-infarct complications and stuff like that, cardiogenic shock.

I thought the attribution approach, which was based on primary care

physicians, was not well suited to infarction, which was hospital based. Defining a single responsible physician I thought was very challenging, particularly relating to high cost items in the cath lab that may have nothing to do with the primary provider. And we had the discussion about that for the last example.

So I found that it was very hard to produce favorable scores on any of these components because the risk piece was missing, and the exact nature of the patient group that we were studying was ambiguous.

I could go through them in detail individually, but I think there's a generic problem here.

CO-CHAIR CURTIS: I think just because most people haven't had the chance to review this type, although many people have reviewed another Ingenix measure, it might be worth just going through your specific concerns about the characterization of the

episode, and what's making you concerned.

DR. MARWICK: Well, in the characterization of the episode, for example, there's a statement about looking at observed and expected costs for CAD and infarction. So we know that that's a heterogeneous group of episodes. It varies from somebody presenting with an acute infarct, transmural infarct requiring attendance to the cath lab, to a non-STEMI, potential medical care in hospital, to complications of all those things ranging all the way down to presentation to a primary care physician.

So this doesn't include just the description of acute MI. It says, "episode results were not readily available for AMI episodes to support a specific analysis of this condition. However, results for CAD and AMI can provide some insight."

So just in the beginning, about the definition of that, I find that very, very broad.

CO-CHAIR CURTIS: Maybe if the developer could comment on that limitation or -- was there a reason you couldn't break out AMI? And this was confusing in my review as well, was is this embedded within the ischemic heart disease measure or is it thought to be distinct? And some of the sample reports that you showed, I couldn't even find AMI in there.

DR. LYNN: It's embedded in ischemic heart disease or coronary artery disease. And with diagnostic-only markers of myocardial infarction.

CO-CHAIR CURTIS: So it's part of a measure that would be recorded but not the entire measure?

DR. LYNN: It's a coronary artery disease reported measure that includes diagnostic evidence of acute myocardial infarction.

CO-CHAIR CURTIS: But what about -- so no patients without that evidence of having experienced an AMI would be included in

this?

DR. LYNN: That's correct.

CO-CHAIR ROSENZWEIG: Is that an acute myocardial infarction in the recent past or ever in the patient's history, or --

DR. LYNN: Well, the diagnosis indicates that it's acute, but --

CO-CHAIR CURTIS: So maybe we could actually go to that.

DR. LYNN: Again, I think this clearly should be -- you know, there's a feeling here that should be event oriented. You know, I think that's a reasonable --

CO-CHAIR CURTIS: Well, did you consider not, you know, doing this a more traditional episode of admission for MI with post -- you know, as a start of the episode?

DR. LYNN: We were asked to provide -- I think we would do it that way. I think we would do it that way, as an event. But that's not --

CO-CHAIR CURTIS: So just to

broaden the discussion and include the other members, if you look at -- if you bring up the spreadsheet that shows the diagnostic category codes, I think it's workbook S-5 something-something, and sheet 4 within it.

CO-CHAIR ROSENZWEIG: Do you know exactly which one would have it?

DR. LYNN: Oh, it should be S-5.

CO-CHAIR CURTIS: So while they're bringing that up, so you know, as the ABMS measure specified in acute admission for an MI, so the 4.10.x1 implies admission for that procedure. This measure, if you scroll down, includes those -- sorry, no, where would it be? Primary diagnosis code, the worksheet, the fourth worksheet in? Yes, I think.

So this is how they're characterizing the codes that are included in the population. And so it's 4.10, and -- but not specified to dot-x1. It doesn't specify an acute episode. So they could be more in the chronic phase, they could be at the acute

phase. So it's a heterogeneous population from that standpoint.

And if you scroll down, do you make the -- they try to account for subendocardial infarction using the specific codes of 410.7x. And then there's the inclusion of the 429.5, 429.6, which is very different than how we traditionally characterize MI patient populations, or acute MI patient populations.

So I think this is probably what you were reacting to. And is -- it's a very different measure. But I --

DR. HWONG: If I could get some clarity around this, especially with the measure developer here. So I guess what the concern is, you know, with these 410 codes which -- or actually not all of them are 410, right, but with acute myocardial infarction, so this could be present on any -- you know, as a primary diagnosis on whether it's a hospital stay or an E&M visit. I see, so

either one could actually be the anchor to start the episode then?

CO-CHAIR CURTIS: As long as it had excluded the clean period, which is think is 30 days here.

DR. LYNN: Yes.

CO-CHAIR CURTIS: So in that case, I think you could have a 30-day clean period, have someone come in to your cohort as a outpatient visit --

DR. HWONG: Yes.

CO-CHAIR CURTIS: -- the code is a 4.10.x2. And they're in your cohort. And that's very different than with very different resource use expectations when someone's not.

DR. HWONG: Got you.

CO-CHAIR CURTIS: So I would -- right, and I think it's two weeks -- two or four weeks after, is how it -- anyway, it's certainly more than the acute admission.

So I think in the -- maybe if I'm hearing the measure developer correctly, that

based on this feedback, would you want to reconsider our consideration of this measure or would you -- I don't know if you --

DR. LYNN: We would have to --

CO-CHAIR CURTIS: -- can do it during this timeframe.

DR. LYNN: Yes, we would have to wholesale change it, which, you know, we'd love the opportunity to do. But I think we've -- well, anyway.

I am -- I think it's reasonable for us to withdraw the measure at this point.

CO-CHAIR ROSENZWEIG: In practice, how -- to what extent are cardiologists and primary care doctors using the granularity of these various individual subcategories --

DR. LYNN: Let me --

CO-CHAIR ROSENZWEIG: -- in cardiology?

DR. LYNN: Yes, in general, we were trying to fit a little bit of a square peg in a round hole here for us. And I -- and

the unit of analysis that's used is the episode of coronary artery disease. It's not used as an episode with AMI in isolation at all, as far as I know. I mean, we tried -- we could test it and we could run it through our data, but the methodology is meant to create an episode of coronary artery disease and mark that there's evidence of acute myocardial infarction as a severity adjustment. And we tried to, you know, configure that to make that meet the call for measures. And I think it was totally unsuccessful.

MS. TURBYVILLE: Just to provide some context, this particular effort, as well as the other NQF efforts, are looking at evaluating measures independently. And as you know, the ETG system -- and Tom, I don't want to speak for you, but it's a system of measures that work together.

And so you know, the -- but we did need to -- we're not evaluating a group, or we're not looking at systems of measures in

this particular effort. We really are evaluating discrete measures. So that's -- you know, they're -- I think that's in response to you trying to fit the square hole in the round --

DR. LYNN: That's correct.

MS. TURBYVILLE: Yes, yes. So to really provide, it was NQF who insisted that these be independent measures and be evaluated independently of each other.

CO-CHAIR CURTIS: So again, I'm not terribly sure how to proceed, this is sort of unprecedented in my experience.

DR. HWONG: I do have a question, just for us to be able to understand the context for the rest of the Ingenix measures, right? But I think there is an Ingenix measure for coronary artery disease, I mean, you know, if I'm not mistaken.

So if this one is sort of modified, like how should we be looking at the next one, right, like in terms of the context?

Like how different --

DR. LYNN: Well, the next one is looking at a disease, a coronary artery disease that occurred for one year, period. It's not event-driven.

DR. MARWICK: Yes, I agree. I think the problem here is that the label here is acute MI. And you know, there we're talking about coronary disease. We do have a little bit of a mirror image problem of capturing the acutes as well. But this is a particular problem, that this is a label of acutes that's being contaminated by other entities.

MS. TURBYVILLE: Yes, I mean it sounds like based on that, that Ingenix, the measure developer, would like to withdraw the measure, at which time we don't have to continue rating the measure and put, you know, the rating through that whole process, as well as the developer.

DR. LYNN: That's correct.

MS. TURBYVILLE: Okay, great.

CO-CHAIR CURTIS: So then we were going to -- right. So in the interest of reviewing one measure from every measure developer while we have Carlos here, for a limited time longer, we were going to go over -- tell me again, Sally?

MS. TURBYVILLE: We're going to jump to the NCQA 1557, the diabetes NCQA measure. Yes, 1557.

CO-CHAIR ROSENZWEIG: You, being the primary reviewer --

DR. HWONG: Yes, I guess so.

MS. TURBYVILLE: Let's give Ben Hamlin from NCQA a few minutes to introduce the measure, and any other approaches that you want to reiterate.

MR. HAMLIN: Thank you very much, Sally.

Is this on? I can't tell. Okay.

Thank you very much. NCQA currently has five condition-based -- the big

five chronics, if you will, total annual measures. So these are a little different from the one you're evaluating right now. So we're looking at total annual cost or resource use for anyone identified with one of these chronic conditions, diabetes being one of the factors.

The measure-eligible populations are aligned with the HEDIS chronic disease manager. So for the diabetes population, the base eligible population looks at a very similar population to that, but is defined in the NQF-endorsed diabetes set, if you will -- I think it's 00623, 0068 or something like that -- instead of NQF-endorsed quality measures.

We only used the RU measure results along with the quality measures, so we felt it was very important to align those two things together. And again, I'll be here to answer any questions you may have about the methodology, but that's really the high-level

overview.

DR. HWONG: Okay. Well, I guess I will start to drive.

In terms of the first area, sort of the measure focus and sort of the importance. So you know, does this address -- sort of 1.A, does this address a national health goal as defined by DHHS or National Priorities Partnership? And I actually rated this as high, I said it seemed to align with National Priorities Partnership for affordable care, elimination of overuse. Also that it again, you know, affects large numbers, there's high resource use, and that there are a lot of societal consequences to poor quality management for diabetes.

So I'll pause there and see if anybody has any other comments in that regard?

(No response.)

DR. HWONG: Okay. Should I move -- should we vote or move on to 1.B? Maybe just go through the importance --

CO-CHAIR ROSENZWEIG: Why don't we go on to the importance one first.

DR. HWONG: I agree. Sure.

CO-CHAIR ROSENZWEIG: Oh, yeah. Why don't we just continue going through the importance ones first --

DR. HWONG: Okay.

CO-CHAIR ROSENZWEIG: -- and then vote on them as a group.

DR. HWONG: Sure. So demonstration -- so 1.B, demonstration of resource costs -- resource user costs, problems and opportunities for improvement.

So what I found here was that NCQA was able to sort of look at their own history in their annual analysis that they've been doing over the last four years in identifying sort of varying resource use or variation in the sort of health services related to diabetes management.

And as such, you know, with that sort of variation, it did seem to imply that

there is opportunity -- you know, opportunity for improvement or modification in that regard.

CO-CHAIR ROSENZWEIG: It should be noted that they're talking about opportunities for improvement in quality of care, mostly --

DR. HWONG: Okay.

CO-CHAIR ROSENZWEIG: -- from their HEDIS measures, as opposed to opportunities of improvement in cost of care or in resource -- specifically resource use. But I think there's a lot of data certainly in there to suggest that resource use --

DR. HWONG: Oh, yeah, I guess it looked like -- it says, "demonstrates substantial variation in health plan resource use for an overall perspective."

So perhaps more of the evidence that's listed is really on quality. But I think it did make mention that, you know, the resource use measures themselves have shown, you know, variability that way.

MR. HAMLIN: One of the things that we had noticed in our annual analysis is that there is a -- they had a variation in resource use between plans achieving the same level of quality, and vice versa. So there's a flaw in the variation on the quality side, aligned with the resource use, and we really have not been able to address specific correlations between those two. So I think there's room for improvements in both areas.

CO-CHAIR ROSENZWEIG: To a certain extent, resource use, I mean a lot of the HEDIS measures are actually performance measures. So for instance, getting an eye exam is a resource use that's a benefit, okay, to the patient, okay? But it also costs a certain amount of money. So there's an overlap -- so a lot of the things you're actually measuring are in that category.

MR. HAMLIN: And we're not specifying that improvement is necessarily lower in resource use. We actually noticed

that there are a few correlations between high resource use and high quality. So we're not saying higher is better and lower is better, we're saying it is what it is.

DR. HWONG: That there is --

CO-CHAIR ROSENZWEIG: Exactly, yeah.

DR. HWONG: Okay, good. So why don't we move on? If there's no further discussion on that, we can go on to 1.C.

I think the purpose of this objective is resource use measure is -- you know, the intent of this is clearly described. You know, I found this was high. I think it described the intents well, the unit analysis is at this regional health plan level. You know, it adjusts the case mix for health plan members and the goal was simply to compare sort of regional health plans versus sort of other peer health plans, it seems like. So I felt like the intent was fairly straightforward and clear. So I would -- I

ranked that as high.

Any comments from the group?

(No response.)

DR. HWONG: Okay. So are we ready -- is this the voting time now?

MS. TURBYVILLE: You have 1.D --

DR. HWONG: Oh, 1.D left, okay.

So the resource use service categories consistent with measure construct. And again, so I looked at this, I felt like, again, sort of the resource use areas, the categories that were listed in terms of, you know, ED visits, hospitalization, procedures, surgeries, you know, pharmacy, et cetera, I felt again that this was fairly comprehensive in terms of the -- you know, the costs that would be generated in terms of management of members with diabetes.

CO-CHAIR ROSENZWEIG: Okay, any questions or comments?

DR. HWONG: I think they'll be up -- oh, sorry.

And the only thing I would mention is, I think there will be time to kind of dive down into a little bit more of the details of what was actually specified. But again, the categories, the broad categories are what you were looking at in terms of generating cost I felt were appropriate.

CO-CHAIR ROSENZWEIG: All right? Any additional comments?

(No response.)

CO-CHAIR ROSENZWEIG: Does the measure developer have any comments they want to add?

MR. HAMLIN: Not so far.

CO-CHAIR ROSENZWEIG: Okay, good.

So why don't we vote on these categories first. Okay, the first one is clearly the -- whether or not it addresses a national goal or priority. And I believe you said it did, it was high?

DR. HWONG: I ranked this as high.

CO-CHAIR ROSENZWEIG: Okay. And

then the second category was opportunity for improvement.

DR. HWONG: Yeah, and I ranked this as high. But like there was demonstrated variation and whatnot, so --

CO-CHAIR ROSENZWEIG: Okay, very good. And then the third category is, the purpose of the objective is -- of the resource measure are clearly described.

Okay. And then the fourth category is that the service categories are included in the research measure consistent with and representative of the conceptual construct represented by the measure, so that it's conceptually coherent.

Someone hasn't voted. Try voting again, maybe you can get counted twice.

Okay, thank you.

All right, so why don't we move on to the scientific acceptability of the measure.

DR. HWONG: Okay. So I guess

we'll spend this portion sort of talking about, you know, again how the measure is constructed.

So I'll talk about sort of a general approach. So in terms of a couple things, sort of the data requirements, they -- you know, NCQA sort of specified that we need demographic data, at least two years of the data. Eligibility, you know, file data. And essentially how it starts off, in terms of this measure, the population -- the denominator population is really individuals who are identified as diabetics, and this can be based on any claim that is paid or unpaid. And there are certain criteria for this.

So when I looked at this, again I think this matches up with sort of NCQA's diabetes quality effectiveness of care measures, so there's this nice synchronization that way. But you know, members can be part of the denominator if they show evidence of, you know, diabetes medications, oral hypo --

you know, hypoglycemic, you know, agents, either in the measurement year or in the previous year. Or alternatively can show two face-to-face meetings, sort of E&M codes in an outpatient setting with any diagnosis of diabetes in any position or category.

So I think -- you know, from there it seems like, again, it lines up with the effectiveness of care measures. It has a way to identify individuals who, you know, have diabetes. You have to show evidence of it in the last year or the current year. And then we can probably move to some exclusions maybe for next, for the conversation.

CO-CHAIR CURTIS: I just want to ask --

DR. HWONG: Yeah, go ahead.

CO-CHAIR CURTIS: -- so on the identification of diabetic on a non-paid claim or a denied claim --

DR. HWONG: Yes, I --

CO-CHAIR CURTIS: -- is that

unusual or is that the standard?

DR. HWONG: It's interesting, you know. I looked at that, and I thought -- so in terms of being -- later on in terms of the calculation, that will be only on paid claims. But I got the -- you know, maybe we can talk to the measure developer, too. But I got the sense that they were trying to essentially scan the entire data set for evidence of this diagnosis code or this usage. You know, uses of medication, put this diagnosis code on a claim, and sort of take that as an individual in the denominator.

MR. HAMLIN: Yes, that is correct. Whenever we were looking for the identification population, we're looking for the diagnosis codes, not necessarily paid claim codes.

And then you're also correct, when we get to the pricing of services on these services, it's only on paid claims or claims that are expected to be paid.

DR. HWONG: Right. So I think if -- you know, I guess that's the idea of trying to be very sensitive, right, out there? If there's any member that has any evidence of some sort of diabetes, you know, diagnosis code, you know, the measure developer has opted to try and include them.

But then for later on, if the costs, you know, in terms of, you know, their costs, if it's all, you know, denied claims, et cetera, then that person would be -- that individual would be excluded. But we can go through sort of the exclusion criteria which might help.

DR. MARWICK: So how did you deal with the person who changes their diagnostic status, the post-operative diabetic who is no longer then a diabetic, or the metabolic syndrome who loses weight and then is no longer a diabetic?

DR. HWONG: I did not see in the specs handling of that. I think it simply

said, you know, if there was evidence of two separate diabetes diagnoses in outpatient setting, for example, you know, those two, if they change -- if clinicians change their mind about the diagnosis and that diagnosis didn't show up for the second half of the measurement year, hypothetically, that person would still be considered diabetic.

I think it's very hard to be able to capture that kind of change. I think this is, you know, just trying to look for some kind of evidence of -- you know, of at least two episodes of coding, you know, in an E&M basis.

CO-CHAIR ROSENZWEIG: I think the general convention is that you don't lose the diagnosis of diabetes. I mean, even with the bariatric surgery or -- as opposed to the metabolic syndrome or the so-called pre-diabetes category, which has other terms attached to it now.

You don't revert to normal,

necessarily, if your glycemic control improves. You're still considered to have diabetes. Now it may be that, later on, a physician might have -- a patient might have a -- a physician at some later time may not include diabetes among the diagnoses, when the physician sees the patient. But in general, it's a matter of basically controlled diabetes.

Now that's different from secondary diabetes, which -- and there are a whole series of codes associated with secondary diabetes. The 249 codes, okay, whereas if a person was like on glucocorticoids or -- and of course, gestational diabetes can revert to normal as well.

But type II and type I diabetes, in general, kind of stick to you.

DR. HWONG: And as you've mentioned, those are actually explicit exclusion criteria, so individuals who -- and

I'll go through some of the specific exclusion criteria. But PCOS, as well as steroid-induced diabetes, if those are coded without evidence of a follow-up -- without evidence of any E&M visit with the concurrent diagnosis of diabetes, then those individuals would be excluded.

So I think there's some -- some thought to, you know, essentially exclude individuals who are being treated for gestational diabetes, exclude individuals who are temporarily being treated for steroid-induced diabetes.

But just to be complete in terms of their exclusion criteria, you know, and this was something that was echoed earlier in the ABMS, you know, measure, I think. But essentially excluding individuals who have any evidence of active cancer, ESRD, organ transplant, HIV-AIDS. I think -- and again, with those special criteria for the gestational diabetes and the steroid-induced

diabetes.

CO-CHAIR ROSENZWEIG: There may be some changes as well. I don't know if you addressed this in the measure, I don't remember if I saw it. But the -- in the past, you know, even if a person did not have a diagnostic -- a diabetes-related diagnostic ICD-9 code, they would be considered to have diabetes as a group if they were on medications for treatment of diabetes.

That's changed, largely because people are using metformin and other agents. A certain percentage of physicians are using them in the pre-diabetes state, as well as for PCOS. So it becomes a little more complicated.

DR. HWONG: Okay, good.

Sorry if I'm jumping around with this. I'm trying to wade through some of these notes here.

But okay, so the only other aspect is to back up a little bit. You know, I

talked about how you would identify these members, how you would exclude certain members. If we just sort of jump back a little bit in terms of the data and sort of what the measure developer has submitted in terms of, you know, what do you do -- you know, what you want to do ensures sort of the integrity of the data that -- I'm trying to think here.

Yes, so I mean, I think there's some mention here that there's no desire to impute or -- you know, with missing -- with any sorts of missing data. Let me see.

Oh, yes, I'm sorry. So -- and again, any of -- as we talked about, the denominator can be defined by anybody whether it's paid or unpaid claims. But the service, in order to be considered for this resource measure, would have to be paid at least in part -- you know, in full or in part by the plan, or that the member absorbed the cost entirely, right. And that this is a service

that's covered under sort of a PMPM payment by a health plan.

So again, just trying to define, you know, if there was some kind of payment or if the health plan is responsible for that cost, that cost would be included in terms of the calculation. So I just wanted to back up on that and sort of cover that topic. Okay.

CO-CHAIR CURTIS: It would be helpful to just sort of refer to what part of the PDF you're in, too, what page number.

DR. HWONG: Oh, sorry.

MS. TURBYVILLE: It's page 9 of the PDF.

DR. HWONG: Thank you.

And let me scroll back here.

Okay, so -- right, and we talked about the exclusions, et cetera, so maybe we can go towards -- yes, and the only other thing in terms of that clinical framework, they do provide a detailed listing in terms of the types of medications that would be used in

that criteria, as well as the specific lists of what qualifies as an E&M visit with the specific CPT codes or, you know, revenue codes.

Okay, so maybe we can scroll down to -- I guess in the printed out version, it's page 13, because I think that takes us to that point. And talk about the sort of comorbid -- comorbid and interactions, you know, in terms of how are we identifying individuals with, you know, certain comorbid conditions.

So the resource -- you know, relative resource use measure is using the HCC relative resource use risk categories. And from what I understand, you know, in looking at this -- looking at this explanation, right, you know, understanding that this is an externally sort of validated risk adjustment method for costs.

Essentially, based on sort of the diagnoses codes, individuals get grouped to one of 184 sort of clinical condition

categories, which ultimately can get rolled up into the HCC, you know, relative resource use categorization, which ultimately has a ranking, right. So a member will be classified, you know, among those. And then the highest ranked -- I guess that's the lowest number, you know, in terms of the system, essentially gets assigned to that patient.

So you know, I think I sort of described that on a high level, and I want to see if the measure developer has any further comment in terms of how this HCCR, you know, relative resource use, you know, categorization is used. My understanding is, based on these codes, you get rolled up and then essentially you get assigned to the most significant -- significant category, unless -- oh, and then unless there's some other code, you know, that is as known interaction which would increase your risk adjustment.

MR. HAMLIN: Right. So you're

exactly right. And there is some interaction with combination code -- HCC codes.

Effectively what happens is, once these categories are assigned, they're also assigned a weight from this. And then an additional demographic and -- so basically age, gender, weight is also assigned. And so effectively what you do is you add up the weight of that member and assign a risk category -- a risk grouping, if you will.

DR. HWONG: Okay.

MR. HAMLIN: So we have 13 risk groupings for each measure.

DR. HWONG: Right.

MR. HAMLIN: So if you take the CC to HCC, if there's multiple HCC's, you take each of those weights. You add the demographic weight, and that's the total weight is what that member is weighted, as their risk overall for comorbidities, demographics, and sort of everything in one bucket.

DR. HWONG: Great.

CO-CHAIR CURTIS: And that's assessed in the 12-month period, or what period upstream?

DR. HWONG: Yes. It seems like that, if I'm not mistaken, it's sort of prior to, you know, not the measurement year, but the year prior to the measurement year?

MR. HAMLIN: No, it can be done in the measurement year or the year prior. So it --

DR. HWONG: Oh, so it's the two years?

MR. HAMLIN: Right.

DR. HWONG: I'm sorry. Okay, so within the two-year period.

CO-CHAIR ROSENZWEIG: Are you using entirely coding data to be able to identify these comorbidities?

MR. HAMLIN: Yes. The -- primarily ICD-9 diagnosis.

CO-CHAIR ROSENZWEIG: All right.

The one potential problem with this that always occurs, and actually I was alluding to it when we were talking about the cardiovascular -- one of the cardiovascular protocols is that there's a notorious lack of use of the complications codes, as well as the -- you know, the 250.xx, there's a notorious lack of use of the codes defining the various complications as well as the codes that identify whether or not the patient has or does not have -- is or is not in good control.

MR. HAMLIN: Right.

CO-CHAIR ROSENZWEIG: So especially among the primary care population. So it becomes a -- it becomes a difficulty in truly identifying the full spectrum of diabetes complications.

MR. HAMLIN: Yes. And we actually -- when we create the HCC tables, which we do, but we don't use every single HCC that CMS publishes, obviously. We try and select -- we go through a process of selecting the ones

that are most relevant to the chronic disease population. But I will agree, there is probably some variation in the area of the use of the codes, where they're not all there. And so therefore, some of the rankings might be -- might have some gaps.

However, in doing this whole idea of identifying at least as many comorbidities as we can, or at least we feel putting them in the right risk groups, which is the one to thirteen, as opposed to just identifying them with a comorbid or no, which is the previous approach that we used, which we feel kind of doesn't really give you that exact same case mix for these patients who have multiple comorbidities, and some are more serious than others. So it's a step in the right direction, it's still not the perfect approach, I don't think.

CO-CHAIR ROSENZWEIG: Yeah, I'm not disagreeing with that. I was just saying that there is a problem with the use of those

codes, which hopefully in the future, physicians will adhere to better, because they'll get paid more.

MR. HAMLIN: ICD10 is going to fix everything. That's my story.

(Laughter.)

CO-CHAIR ROSENZWEIG: But the HCC process has been validated with respect to costs, hasn't it?

MR. HAMLIN: Yes.

CO-CHAIR ROSENZWEIG: Yes.

MR. HAMLIN: Yes, it has.

CO-CHAIR CURTIS: I just want to follow up on that. So if you're using the comorbidities for risk adjustment and you're assessing it in the measurement year, isn't that potentially explaining away some of the differences or the variation that you're seeing?

It's different than what we do for sort of the outcomes measures?

MR. HAMLIN: So when these get

reported then, we actually back up a bit and then report them out by the risk group, and also by age and gender categories. So while we roll them -- we take these things into account in the risk adjustment, in order to do calculations. The reporting out then is done in these member cohorts, which are the age, gender. So for example, in diabetes, one of our member cohorts would be males 18 to 44, HCC category 1; males 18 to 44, HCC category 2 would be the second one.

So they get rolled up into groups with the calculation process. But then when they're reported out, we actually do report them out in sort of an expanded set in these different cohorts. So we do -- while some of it's adjusted away, we then do try and identify them in the reports specifically by these member cohorts, which we feel are most applicable to this condition.

It's a little confusing, I fully understand, but it's --

DR. HWONG: So just - I'm sorry, I want to understand your concerns so that, like, in contrast, previously the risk adjustment was done on data prior to the measurement, right, like that one other example? So you're concerned -- and I want to understand your concern, but that if there is data taken to understand their comorbid status during the measurement year, is that a problem or --

CO-CHAIR CURTIS: I'm trying to unravel it in my head, but it just seems like it's not intuitive or it's just, again, different, setting a red flag. And maybe that's appropriate for costing, I don't know.

But for instance, if you have increased resource utilization with increased use of the diabetes complications codes, because you're seeing the patients more frequently, then you're adjusting that away because they have more diabetes complications coded.

DR. HWONG: Right.

CO-CHAIR CURTIS: It just seems like it's circular.

DR. HWONG: Yes.

MR. HAMLIN: I should clarify too that the weights are actually calculated on the previous year's data, because that's how we do -- we do the calculation of weights based on prior year. The identification of each person for the HCC is --

DR. HWONG: Yeah, that helps.

MR. HAMLIN: -- done in the measurement year itself. So there is a one-year lag. But they're updated every year, so we are using hopefully the -- you know, the most current available data for the calculation of risk adjustment.

DR. MARWICK: What does the risk adjustment predict?

MR. HAMLIN: Well, I'm not sure what it really predicts, but I think what it does is it really balances out some of the

factors that allow us to not create comparable populations for the plan. I mean, we end up ranking plans, and so the idea is to create comparable populations for our approach.

So what we do is try and adjust away the mitigating factors that would really skew the plan-specific populations one direction or the other. Beyond that, I'd have to probably get someone who is more technical in explaining the specifics of that.

CO-CHAIR ROSENZWEIG: But was it designed for severity of illness or was it designed to predict costs?

MR. HAMLIN: Well, a little bit of both. It really was more on the -- the ones we selected were the ones that were most predictive of costs for this population. And again, we're looking sort of commercial, Medicare, Medicaid plan populations. And we take those factors into account when selecting and designing the HCC-RRU tables that we used to do the risk adjustment. So they're the

ones that are the most predictive of cost.

MS. CLARK: They were designed by -- it was for Medicare. It's in use right now by Medicare for paying managed care plans, their monthly capitation rates or risk adjusted capitation rates. So they're looking at, you know, the patients that these managed care plans are getting enrolled, and how predicting their costs. So it is for predicting costs.

I think the scores that are generated, you know, if you have a value of one, if some -- if a patient has a value of three, then they're three times more costly than the average.

CO-CHAIR ROSENZWEIG: Yeah, I know it was used for the Medicare health support program, when they put that together.

DR. HWONG: Great. So sort of moving along. We figured out, you know, who's going to be in this measure, right? And in terms of the cost, we know the resource

categories. And so I think maybe here we can, you know, start to introduce what is going on in terms of the cost data, right?

Here is where the measure developers introduce sort of the standardized price tables. So this is not going to be reflective of any health plans, actual contracted, you know, contracted fees and rates. It's not about charges or costs, this is just really sort of normalized -- or standardized, rather, to -- you know, across all, you know, health plans or across the data in terms of the standardized pricing tables.

So every service that is associated with these members gets mapped to, you know, this table, right, in terms of the prices there or the costs there, excuse me.

So the advantage of that, in terms of trying to make, you know, sort of comparisons across health plans, you know, that sort of essentially, you know, with everybody in terms of the same sort of E&M

code or whatever, would just get mapped to sort of the same costs. And so I think it allows for a greater comparability or comparability between health plans, ultimately, which is kind of the intent of this measure.

So the one thing, when I was thinking about this, right, where this is not so much about sort of the costs of care regarded to an episode or truly the management, per se, of diabetes or of diabetes patients. I got the sense that any type of inpatient admission, any type -- I mean, if someone, God forbid, had a horrible motor vehicle accident and, you know, was laid up in the ICU, those costs would actually still be associated with, you know, these members who have diabetes.

So it's really -- you know, I mean, and I'd love to hear from -- I'm just looking at the measure developer here, but yeah, it's interesting that, you know, in some

ways you're very specific about who these people are, these diabetics, but there's -- you're really sort of taking this global cost of, you know, sort of cost of care for this diabetic regardless. I mean, I know there's some risk adjustment, regardless of, you know, what the management might before, which could be completely unrelated.

MR. HAMLIN: Well, it is and it isn't. I mean, yes, you're right. This is a great annual snapshot of the utilization of a member with diabetes. And that really is truly what the measure is designed to do.

But my best example I can give, you know, is how do you know the person didn't fall over and break their arm because they had a severe hypoglycemic episode?

So again, we don't want to do that identification of specific episodes to the chronic condition itself, we want to give them a total resource use snapshot, broken down by, you know, 21 service categories for that

member and what that member or that population might look like, given that perspective.

So we don't make any associations between specific things. We do exclude, as you mentioned earlier, the sort of high cost conditions, HIV, active cancer, transplantation, because we feel that those -- you know, a small percentage of the population might skew the costing approach too much for a specific plan.

But we really feel that it is important to capture all service utilization for that member with diabetes over the year, whether that's directly attributable or not to the diabetes itself.

DR. HWONG: And so another question that I had, when I was thinking about this measure, was you know, so if we're capturing global costs, right? I mean, I'll just say global for now, but like, you know, for these diabetics.

And then they have these

comorbidities and they sort of happen to fall into the heart failure category, et cetera, you know, how do you sort of distinguish for a member -- you know, so the member essentially gets double, triple counted, you know. When you roll it up in terms of a health plan, how do you sort of separate that out where -- because it's not like specific services, you know, are attributed to the episode.

It's just globally the cost of someone who has -- you know, a member who has these multiple comorbidities end up -- you know, end up sort of showing up in terms of costs that are used for a health plan multiple times?

MR. HAMLIN: So I should also probably qualify it. We actually only are able to price around 82 percent of the actual costs. So we have tested each one of these, the coding structures in a variety of health plans. We have a large research database, is

what we call it, that helps us to pilot some of these costing structures through this.

We have to look for the reliability of the paid claims, and have to be sort of not -- you know, we have to look for duplicate claims and things like that. And there are certain services we just can't price because they're just too unreliable.

So we're about 82 percent right now with those, and so we're only pricing select services, but again it's about 80 to 82 percent of those associated costs. But once a person's been identified with diabetes, any of their utilization that they incur over that measurement year is attributable to that person sort of being rolled up in these categories.

If there -- if the service -- if the code is in the standard pricing tables, it should be counted towards that member for that year.

DR. HWONG: Okay.

MR. HAMLIN: That's basically the way we look at it.

DR. WEINTRAUB: So this is the opposite choice that we saw from the -- from this morning, from the acute MI, where there was a detailed attempt to figure out which codes to attribute to the MI versus everything you see. This is actually -- this is a simpler approach. And you're not trapped into trying to figure out what's in and what's out and you don't have to -- the problems about updating, you just count them all up and multiply.

So it's simple. I like it better. That's what people usually do in cost effectiveness analysis. Its downside is also obvious, that you introduce some noise that may make it more difficult to really distinguish between providers.

But again, I think you've got the point exactly right here. Someone falls down and breaks their arm, does it have nothing to

do with diabetes? You really can't say.

CO-CHAIR ROSENZWEIG: The data also indicates that with almost -- with a wide variety of surgical procedures and hospitalizations that are unrelated to diabetes, if you're in the hospital, you tend to be in the hospital for at least a day longer if you have diabetes. So length of stays are -- tend to be for almost any condition, cholecystectomy, for example, approximately one day longer, on average.

DR. HWONG: And I was thinking like this rolled-up method, right, where it's not really about the whole episode of management, you know, of the diabetic patient, per se. But this is, you know, probably okay in terms of useful and aggregate on a plan level, I think this would be more problematic, you know, going down to that individual, you know, physician attribution level.

Again, where you know -- you know, I come from the perspective where we have had

to actually implement measures in terms of quality profiling, et cetera, and I guarantee you, as some of the things that, you know, physicians would come back with which is, you know, this is unrelated, you know, I have -- you know, my patients happen to be X profession -- you know, whatever, and these sorts of issues.

So I think, you know, that sort of global concept I think does, actually in general, fit better for a larger sort of unit of analysis.

MR. HAMLIN: The approach works in the health plan in large physician group level. It does not really work at the individual provider level. You have to have a sample size of at least 400 or so patients with the disease, which I think is unavailable to many physician groups.

DR. HWONG: Yeah.

MR. HAMLIN: It works in those scenarios and ACO's, and whatever else you

want to call them. But I mean, it's got to have a fairly decent-sized population in order to use with the data that encompasses all the different systems.

CO-CHAIR CURTIS: Just remind me then, where is this plan to be attributed to? What level?

MS. TURBYVILLE: That's 11.3, page 26 of the PDF is where you'll find what the measure developer selected as the --

CO-CHAIR CURTIS: We'll get there. We haven't passed it yet. Okay.

DR. HWONG: Yeah, but it looks like it's broken out by product -- I'm sorry, yeah, product line. So it will be on health plan, but they'll segment the populations like Medicaid, Medicare, commercial.

And then I think you're mentioning that it would be large provider groups. But I didn't actually see -- maybe I didn't see that.

MR. HAMLIN: It's been tested and

used in that environment. We don't actually maintain that group, we're using different health plans right now. So it's --

DR. HWONG: And so I had sort of assumed that this whole thing was about essentially on the health plan level. But perhaps segmented by line of business.

MR. HAMLIN: Right.

DR. HWONG: Right.

MR. HAMLIN: It's commercial, Medicare, Medicaid and we divide them by HMO, PPO and further. And so it's --

DR. HWONG: Okay, great.

MS. CLARK: I just had one question, in terms of the standardized cost calculation. And so that, again, is based on the claims data? You're calculating standardized -

MR. HAMLIN: Primarily the Medicare fee schedule and there's a lot of adjustment for the commercial data that is maintained in our data system. I think they

use an RBBS to adjust the Medicare fee schedule primarily, but we use the national Medicare fee schedule, I think, as the base for the standardized costing. Because there's so much variation across the country in prices actually paid, contract by contract, and we don't really want to dive into getting each individual contract price and trying to make some kind of adjustment for it.

DR. HWONG: Right. And I think also for the pharmacy costs then, you're sort of using sort of the -- was it average warehouse -- you know, the AWP pricing, right?

MR. HAMLIN: Right.

DR. HWONG: Which -- I mean, which may not in reality sort of reflect again contracting with PBM's and sort of how certain things will become different tiers in terms of cost. But you just sort of use this standard AWP, which you know, can over-estimate, I think, you know, some of the costs there.

All right. So the one other thing

I thought was interesting was, in terms of the cost calculations, they actually -- the measure developer spent some time to pull out the hierarchy in terms of cardiac catheterization procedures, coupled with CABG and how to sort of separate that out. Where everything else -- everything else in terms of this whole thing you just throw into the bucket, right? But for -- specifically for like cath or invasive -- like there's this strict hierarchy where you'll sort of throw out some of this and sort of take -- it seems like to take the most invasive, you know, cost and sort of -- and use that alone. So if someone has, you know, a cath go in, you'd understand or whatever. Then a CABG, it looks like, you know, you just --

MR. HAMLIN: So those are actually included in the cost component as well.

DR. HWONG: Okay.

MR. HAMLIN: The service frequency component is an additional component of sort

of what we call select procedures that are relevant to that condition. It's primarily in diabetes and cardiovascular, which are the two measures under evaluation.

So we look at frequent procedures that are attributable to populations so that you can look at the standardized cost. We have the surgery and procedures both in inpatient and outpatient level, but then you can also see on a per unit per year basis, the frequency of the procedures performed in this population. So you're looking at the frequency of CABG's in the diabetic population for this measurement year.

So there's a frequency number and there's a cost category. And actually the CABG's are included in the cost component of the total roll-up. So you can sort of cross compare those two things, and hopefully that will give you more information.

DR. WEINTRAUB: Just to clarify, you're not using different methodology to

cost?

MR. HAMLIN: No.

DR. WEINTRAUB: Okay, thanks.

MR. HAMLIN: No. It's -- since you're counting the year, you're mapping it for the costs, you also put a checkmark on the frequency category, effectively.

CO-CHAIR CURTIS: How did you define the procedures that were relevant? Obviously cath, PCI, CABG makes sense. But why not amputations or progression end state renal disease or other diabetic-relevant complications?

MR. HAMLIN: Many were a derivative of a past HEDIS measure that was frequency of selected procedures. This went through several expert panel reviews, and they sort of decided this was a good list to look at the procedures that perhaps were either in the appropriateness over-use category, as future thought into measurement in those arenas.

They were also felt to be relevant to this population so it would be of interest to -- you know, specifically of interest to plans who were trying to identify opportunities to improve in certain areas, primarily with procedures that are frequently performed.

And they can look at -- and again, you can -- when the plans compare each other to their peer group, they can look at these frequencies and compare themselves to other frequencies that are displayed by other plans, to see how they compare, again, across these different categories.

CO-CHAIR ROSENZWEIG: I didn't see pregnancy list in the cost calculations. Are you -- you're not excluding women with diabetes of child-bearing age, are you?

MR. HAMLIN: Well, gestational diabetes is an exclusion from --

CO-CHAIR ROSENZWEIG: No, I'm talking about people with diabetes who become

pregnant.

MR. HAMLIN: On maternity, I believe there's a series of maternity codes in the cost calculation tables. I couldn't tell you which ones exactly. There are some that are not included, however, and I don't know exactly -- I have to get more information on that for you. I know that maternity is one of those difficult areas we're struggling with right now in identifying what should count and what should not count.

Certain maternity codes are in the 82 percent that I mentioned, but there's also a series that are kind of in that gray area that they're not consistent -- not consistent enough so that we can't price them accurately, if that makes sense.

CO-CHAIR ROSENZWEIG: Okay. But they do represent a very significant cost component?

MR. HAMLIN: Yes. And I know a number of them are included, I just couldn't

tell you what percentage of them -- you know, which codes and what percentage of the services are included in the current 80 versus the 20 percent of services that are captured and priced.

CO-CHAIR ROSENZWEIG: Okay, thank you.

DR. HWONG: All right. So I think we have moved through sort of -- what is it, 9.7. I think we're sort of moving on if people are following along in the sheets, let's move to sort of page 19, all right. And we'll go through a couple of these other categories then.

So care setting provides information in which care setting is encompassed. So again, since this is the whole kitchen sink in many aspects, not in a, you know, direct -- not to be negative, but basically a lot of these areas, you know, it is covering some ambulatory care, inpatient care, laboratory, pharmacy, you know. It is

covered, you know, within this measure.

Going to sort of item 10.1, sort of the risk adjustment method, I think we've spent, you know, some time already sort of discussing the details of that. And sort of how that is organized. I think in the end it sounds like, you know, when you're comparing health plans, you're sort of able to stratify, you know, in terms of your reporting in terms of along these. So in these particular cohorts, or are these like categories?

MR. HAMLIN: We're actually doing a comparison, we actually do the roll-up and then do the comparison. But then the data is available, if you keep clicking down through the published reports of -- by these individual member cohorts to the strata.

DR. HWONG: Okay. That's great. I'll pause --

CO-CHAIR ROSENZWEIG: Is your data -- you need 400 patients, distinct patients with diabetes to compare two different groups.

Is that with the risk adjustment included?

MR. HAMLIN: Yes. You have to have a minimum population of 400 diabetics in your plan, and then everything else gets -- it's not an age cohort, it's totally population. In diabetes, the populations in reality are much, much larger than that. So it's -- we rarely exclude any plans because the minimum population is such a size.

CO-CHAIR ROSENZWEIG: But for physician groups and for individual physicians, it becomes a real problem then?

MR. HAMLIN: Physician groups it does. We're actually seriously considering dropping that minimum sample size down to probably about 250, which may become more attainable for physician groups. I think individual physicians still probably may be a problem because of the -- this HCC methodology has shown that it actually has the same level of specificity at the lower population sample size.

We just -- we're staying with the 400 because it's been tested, it's been run over several years, we're very comfortable with that. Right now the plans seem able to meet that goal, there are very few that are limited from diabetes. So we're sort of holding to something that's constant and, you know, updating other things at the time because they're very complex measures, we don't want to overload the plans with a whole series of changes every year. So we're sticking with the larger populations because it seems to work in the plans.

As we think about -- as we work with groups like IHA to test these with physician groups, and with ACO's coming out, we'll be reevaluating those criteria and retesting different samples, minimum sample sizes. But for now, like I said, it -- the 400 seems to be perfectly appropriate for the health plan population, because almost every plan can achieve that for diabetes.

DR. HWONG: Good. It's perfect. So I'm trying to keep us on the items here.

So any other questions in terms of the risk adjustment methodology? I think we're on -- sort of moving down to sort of 10.2. Long list, we're almost there.

The stratification method, we talked a little bit about, again, sort of health plan and product line. The other thing to mention here is that it seems that you have sort of the resource use categories like you actually will break it out so you can compare across health plans. Sort of inpatient utilization and ambulatory, pharmacy, et cetera, so you can kind of break out and see sort of which areas, you know, may be -- you know, the plans are sort of variable. So I thought that was, you know, again a nice -- nice way to be able to slice it up in terms of stratifying this group and their reporting.

Anybody have questions further on stratification?

(No response.)

DR. HWONG: Okay. All right, and I think we've spent some time on the costing method, again, with these standardized pricing tables. There is a large -- you know, there's a lot of detail here in terms of how to do this for sort of inpatient facility, you know, services, you know, length of days, et cetera.

You know, I -- maybe what I could do, if the measure developer -- you know, if it would be helpful for the group also on, you know, maybe if you want to go on sort of maybe on a high level? I know it's hard, because there's multiple sort of categories --

MR. HAMLIN: Right.

MR. HAMLIN: -- and there is, again, you know, I look at this and, you know, was trying to remember sort of a length of stay, take a length of stay and then you multiply this. And so I mean, I wonder if there is some way to kind of capture this, if you could, you know, for just the group?

MR. HAMLIN: Well I mean, again, it effectively -- you sort of capture the appropriate level services, and then you map those individual codes to -- you know, to the standard price for the inpatient. I mean, the reason we provide specific steps for the different categories is because with inpatient, obviously you have length of stay and issues that we -- a per diem multiplier that we use for length of stay. Again, for the category, it's longer lengths of stay.

You know, the outpatient are generally fairly easy, it's kind of an ICD-9 code and map -- or CPT map for services, but very specific pricing. So again, we sort of provide individual steps for each service category because then the mapping is slightly different, but there are certain considerations, primarily on the inpatient side. And that's the high level, I don't know how much detail you want me to go into.

It's very long and extensive, but

again, to create consistent comparable populations, we want to make sure each plan is doing it absolutely correctly, each -- and absolutely the same each time.

DR. HWONG: Right. And one other thing, and I was trying to understand this, and maybe I have missed this in the details. But in terms of the outlier, potential outlier cost, when you're doing the costing and then, of course, the services, and when you look, you know, overall for any given number, what is the -- you know, I know some of the other measures have used Winsorization technique, you know, to just chop off the ends at the 99th percentile, you know, and out.

But what is the technique for this?

MR. HAMLIN: So we -- all plans that report their data to NCQA, their observed data to NCQA --

DR. HWONG: Right.

MR. HAMLIN: -- goes into a bucket

and gets calculated. We calculate expecteds. We report out any plan that is between .3 and 3.0, so we -- if I could just back up a little bit.

DR. HWONG: Okay.

MR. HAMLIN: We calculate an observed to expected ratio for each plan, specific to that plan. We then do a normalization process where we then take all the plans and normalize their means to one in order to create comparable plan populations.

So then we have this nice sort of mean of one, there are some plans above and some plans below. Plans that fall outside of the .3 to 3.0 range are considered outliers, and are not included in those calculations. So we don't -- we sort of narrow the field a little bit there.

We do actually Winsorize because we actually Winsorize -- when we display these results, we actually do Winsorize any outlying plans down to about -- what is it .5 and 1.5

because they don't fit in the display graph. So there is sort of a two-step of the calculations. Anyone who falls outside of the .3 or 3.0 gets eliminated from the calculations because they're considered to be an outlier.

DR. HWONG: Okay.

MR. HAMLIN: It's a very small percentage right now, less than one percent -- less than .1 percent, I'm sorry, of plans reporting.

And then again, for the display purposes, for reporting these measures, we actually Winsorize any plan between, you know, .3 and .5 to the .5 ratio. We have a little special designated symbol that they get so they show up as a Winsorized plan versus dots on our graph. I provided you a sample display of that.

DR. HWONG: Okay. Great to know that, that's helpful.

Go ahead, yes. No?

Anybody have any other questions?

MR. HAMLIN: Feel free to correct me, Sally, if I'm --

CO-CHAIR CURTIS: Just one clarification, I'd like -- if it's within a calendar year, let's say someone gets hospitalized on December 28th and they're in the hospital. Do you only account for those days within the calendar year?

MR. HAMLIN: Within the calendar year, yes.

CO-CHAIR ROSENZWEIG: Are you able to collect any socioeconomic data?

MR. HAMLIN: Right now we only get aggregate plan level information. We have continually tested member level data in the health plans and the things that are highly inconsistent are race and the socioeconomic status. Gender is fairly reliable, age is fairly reliable when everything else is across the board from two percent to 98 percent.

Some plans are actually not

collecting that data now, purposefully, because they feel it's a liability. So we're not able to do any kind of reporting out of that, by those stratuses, at least, which is why we restrict it to age and gender. The data's just not in the plan systems, and they won't give it to us, even if it was.

CO-CHAIR ROSENZWEIG: What about duration of diabetes?

MR. HAMLIN: That would actually be really interesting. I think that would be something that would be really interesting for us to try and look at. But I don't -- we don't currently collect that. So again, we just receive aggregate level population data from the plan of all their observed. We don't get any individual level member data. It's very hard to report.

CO-CHAIR ROSENZWEIG: I'm not suggesting you add it.

MR. HAMLIN: You know, always open for suggestions.

(Laughter.)

CO-CHAIR ROSENZWEIG: Thank you.

DR. HWONG: Okay we're making progress here.

So if we move on to, again, the -- let's move down to 11.5 -- we're almost there. So subset requirements, I think it's already been stated that, you know, you need have to have a population of at least 400 observations for you to be, you know, included in the measure, which seems to make sense. Especially, you know, essentially they've done sort of an analyses on this, you know, in terms of their observed to expected ratios.

Okay. Benchmarking, right, so 11.6. So again, you know, from what we've heard so far, essentially, you know, this would get again normalized versus sort of the average in there would be sort of one. So you'll calculate sort of these ratios. And I think you can then, with that, sort of, you know, see how far, you know, any given health

plan sort of deviates from that in sort of a positive, more resource use intensive versus less resource use intensive, you know, greater than one, less than one, and sort of see, you know, how many standard deviations, you know, a health plan is out. So there's a way to kind of, you know, take a look at that.

MR. HAMLIN: And we do this annually, so there's no way right now we could trim the data. We don't have the capacity data-wise, because the number of data points required for each of these measures to do that. We're hoping to do that in the future, but right now it's an annual snapshot, and you can, in fact, see for that year how far away you were from the mean, so to speak. But that's really about the best we can do at this point.

DR. HWONG: Okay. All right, any more thoughts on that?

(No response.)

DR. HWONG: Okay. So why don't we

-- actually, I think we can -- should we pause here? Because I think next we'll dive into sort of reliability testing and validity.

CO-CHAIR ROSENZWEIG: Why don't we vote on these --

DR. HWONG: Yeah, I didn't know if we wanted to, yes, since we spent all that time on 2.A.1 or shall we do that all in --

MS. TURBYVILLE: If you want to vote on 2.A.1 now, you can. And then you can move into reliability and validity.

DR. HWONG: Yeah, why don't we sort of have a sense of progress, right, you know, before all of that sort of discussion and detail.

So let's see -- so for 2.A.1, I mean, given -- so having gone through all this sort of sub-subcriteria, you know, I felt like this measure was, you know, very well defined, very precise. And so you know, in general, I looked at this and I thought, you know, there's been, you know, a lot of consideration

for different scenarios and factors and sort of I felt like this was -- you know, I ranked this as high, in terms of the 2.A.1. But let's see how the voting turns out.

Any other comments then, before we go to --

CO-CHAIR ROSENZWEIG: Just the one comment that I had already mentioned, is that the traditional way of identifying patients with diabetes by medication is going to become more and more of a problem in the future. So I don't disagree with what they have been using in the traditional method. But the fact is that you're going to be seeing more and more patients on certain medications that were used for diabetes that don't have diabetes.

DR. HWONG: Right. And I think I saw it --

CO-CHAIR ROSENZWEIG: It's probably not statistically an important issue, but it probably will be in the future.

DR. HWONG: Right. And I did see

that they had already taken -- taken that into account to some extent. I mean, I think they're going to have to continue in the future, but you know, for example, metformin alone cannot be used as one of the medications, I think, to identify someone as being diabetic. It has to be sort of a combination with another class of diabetic drugs because metformin can be used in -- again, for these other applications like PCOS and whatnot, right? If I'm not mistaken?

CO-CHAIR ROSENZWEIG: That's right. But there was a paper that came out showing that the alpha-glucosides inhibitors can also prevent diabetes. So --

DR. HWONG: Right, yeah, so this -- I think that's a point well taken.

MR. HAMLIN: Metformin is a constant thorn in the side of our diabetes measures, both on the EOC side and on the RE side. So there's other things we're looking at. I mean, again, we sort of deferred to the

endorsed diabetic tested and approved identification algorithm we've used for a number of years in the HEDIS population, I think, to keep this population. But I agree there may be some adjustments. I mean, we analyze the medications and the codes annually for inclusion in the measure, and any reflection on the diabetes side will be reflected in the RE measure as well.

CO-CHAIR ROSENZWEIG: Yeah, I'm not disagreeing with you, it's just the fact that it has to, in fact, for your harmonization purposes, you have to have the same methods for identifying patients with diabetes for your quality measures, if you want to coordinate them with your cost -- with your -- excuse me, resource use measure.

DR. HWONG: Okay, great. So why don't we open it up for voting right now?

CO-CHAIR ROSENZWEIG: Yes. So for 2.A?

DR. HWONG: Yes, 2.A.

CO-CHAIR ROSENZWEIG: 2.A.1.

DR. HWONG: So again, you know, is this measure well defined, precisely specified so that this could be implemented across and used -- you know, have results that allow for a good comparability?

CO-CHAIR ROSENZWEIG: All right.

DR. HWONG: All right.

CO-CHAIR ROSENZWEIG: And 2.B.1?

DR. HWONG: Yes. So 2.B.1, all right, measure specs are consistent, you know, with evidence presented to support the focus of measurement.

You know, so this one was a -- you know, in terms of my other comments about, you know, is this really episode -- you know, diabetes management? I think this kind of comment comes in to this category, right? You know in terms of the target population. I think this is where, you know, I felt like again it's large sort of global resource use, and that it would be very interesting to be

able to sort of narrow that down with additional thought. Like some of the other measure developers have gone through that step to say, you know, what is actually really associated, you know, with diabetes management.

So you know, I can kind of see how that could be, you know, useful in that regard, you know. I understand there is that concept of, you know, if someone has a hypoglycemic event and injures their arm, you know, but I do still see how there's a lot that, you know, it may be telling you more about sort of the, again, services as opposed to, you know, broader, as opposed to something that's really sort of diabetes management focused.

So I felt that it was actually more of a moderate for myself.

CO-CHAIR ROSENZWEIG: Any other comments?

(No response.)

DR. HWONG: Okay. So why don't we move on to -- I think we're going to get into sort of validity -- validity testing, right?

CO-CHAIR ROSENZWEIG: Correct.

DR. HWONG: And I would love to take advantage of having Carlos here at the table here. But so when I looked at, again, sort of validity testing demonstrated that the measure data elements are correct and the scores reflect -- you know, correctly reflect the cost of care for resources provided. I know there was this sort of large document included about this 2005 study that, you know, tested a lot of it on sort of face validity, right, you know, looking at it and having these extra panelists take a look and say, you know, member costs are actually high.

You know, with this calculation we see that, you know, costs are higher for AMI patients, you know, heart failure patients, lowest for asthma. I mean, there was sort of this iterative process where, you know,

clinician group or advisory group got to look at that.

I think where I was trying to figure out, in terms of the validity testing was, you know, again like, you know, we're using this sort of standardized pricing table, which I think has some advantages, it will allow for comparability. I just didn't know in terms of, you know, how valid that might be for, you know, some, whether it's plans or groups that have got some sort of things set up in terms of their, you know, how you would actually calculate their costs. I didn't know if that was, you know, something that could be looked at, you know, in terms of this pricing table versus if you were to actually use, you know, costs, you know, per actual service.

But maybe I could also turn it over to Carlos, you know, if you have any comments like here at all. Not to put you on the spot, Carlos.

MR. ALZOLA: I don't have any real

specific comments as to compare standard costs versus actual costs. Clearly you are getting the variability, a lot of the variability out of the calculations when you're using standard costs. But if the point here is to compare plans, and you seem -- it seems to me that using standard costs is helpful because everything -- all those regional variations that could occur because of -- or because of difference in contracting are not really reflecting of quality or efficiency of care. It's more the ability to negotiate a contract than you see done there than a reflection of real costs.

So I think using standard costs is a good approach.

DR. HWONG: All right.

MR. HAMLIN: One of the things that we actually really focused on, now that we have a really well defined methodology, it gave the risk adjustment in the service category that we're comfortable with.

And the things that we're working on now, just FYI, is that we're looking at perhaps a measure of here's your standard cost and here's sort of a delta measure. The difference between irrational cost and standard cost. Actual costs are just a political hot button. They will not give us actual costs. So we were trying to -- and total expenditures is another one we're looking at.

Using, again, same methodology for the same population, but you know, can you give us the difference, if you will, in somehow using that. But those are -- we had to stay away from the actual costs and use the standard costs. Not only that, but we could do that on a national scale. But there is a lot of aggregate variation. So there is some weakness in certain areas, probably, where these costs may not be reflective of actual costs paid in that market.

DR. HWONG: Sure.

DR. WEINTRAUB: These standardized costs are payments, is that right?

MR. HAMLIN: Yeah, it's based on the Medicare fee schedule and it's sort of -- yeah. And it's a national average so, you know, there's some adjustment. But --

DR. WEINTRAUB: Because the Medicare fee schedule, there are both fees and payments.

MR. HAMLIN: Right.

DR. WEINTRAUB: So which is it you can use? The fees are regionally based, rather than national as I understand it.

MR. HAMLIN: They are. But we usually use it on national level.

DR. WEINTRAUB: So you do some sort of averaging of the fees, just come up with a national --

MR. HAMLIN: We don't have a national standard pricing tables. I mean, we don't have regional standard pricing tables, we have just the national standard pricing

tables. So it's kind of a -- it's basically an averaging of the fees for each of these individual service units, or whatever you want to call them.

DR. WEINTRAUB: So again, there is societal costs is a construct, but not something you actually ever get, right? So we -- you know, you pick your poison and hope you can live with it.

CO-CHAIR ROSENZWEIG: I assume you can't take into account the costs of co-pays or whether or not a particular service is covered?

MR. HAMLIN: We think that some of the utilization patterns are very reflective of benefit design, but we can't capture those things at this point in this methodology. It's just too difficult. Again, there's already 10,000 data elements per measure, you know, so expanding that out to capture additional ones, it would just be overwhelming.

DR. HWONG: Okay. Maybe we can move on then to -- let's take a look here. So it looks like we're down to -- oh yes, this item, 3.1, sort of describe the impact of exclusions, transparent in criteria. Am I moving along appropriately? Am I missing something here?

MR. AMIN: Quick process question. Did the group decide not to vote after 2.B? Maybe we should vote at 2.B.1 and then 2.B.2 before we get too far down?

DR. HWONG: No, I think that's good. I'm sorry, yeah.

CO-CHAIR ROSENZWEIG: We did vote at 2.B.1, didn't we?

DR. HWONG: We did 2.B.1, but the 2.B.2 --

MR. AMIN: Oh sorry, 2.A.2, correct. So it's 2.A.2.

MS. TURBYVILLE: 2.A.2 was skipped inadvertently.

DR. HWONG: I'm sorry, 2.A.2.

Sorry. There you go. Okay.

MS. TURBYVILLE: I was waiting for a natural break in the conversation.

MR. AMIN: There's 2.A.1 and 2.B.1. Where's 2.A.2?

DR. HWONG: Reliability testing.

MS. TURBYVILLE: Page 29.

MR. AMIN: Page 29?

MS. TURBYVILLE: Yes.

MR. AMIN: So it is, but it was further down on the list. Okay. Yeah, it's ten pages later.

DR. HWONG: All right.

MR. AMIN: So it's not my fault.

(Laughter.)

DR. HWONG: All right, so maybe I can try and address -- yeah, I'm trying to like it. Hard for me to keep track of this, too.

So okay.

MR. AMIN: Okay.

DR. HWONG: So if I can address a

little bit on this, the reliability testing, when I looked at this.

And so again, this is sort of saying, are the results repeatable, producing the same results a hyper portion of the time? And again, so the advantage here for the NCQA, this measure's been in use for the last like four years or more.

And then -- and so that, you know, annually -- again, you can't -- it sounds like it's not -- I think you made the caveat earlier, it's not sort of a trending, you know, not used for trending purposes, right, but in general that the plan measurements appear sort of stable over time, over these four years, right? That nearly 90 percent of the plans shifted, at most, sort of one quartile within their ranking? And that you know, there was a significant number that -- I'm not going to go down and look at the exact statistic, but did not actually change quartiles at all, so that there is this sort

of stability from year to year measurement. Which I think is very -- I mean, you know, obviously this is a criteria that's important, right?

There's nothing more frustrating for, I think, whether it's physicians being graded or health plans, et cetera, that, you know, one year you're top 25, next year you're bottom 25, right. Then you sort of flip flop back and forth.

So the fact that -- you know, the advantage of this is having sort of this experience over the last four years, showing that again, you know, the vast majority do not move sort of significantly from top to bottom, back and forth, I think is comforting.

CO-CHAIR ROSENZWEIG: Is there anything in your methodology that you're proposing here that's different from what you're currently reporting?

MR. HAMLIN: No. The only thing that's going to be -- every year we expand the

service category. So the data we received last year had I think two fewer service categories. This year I'll have those in there, and then next year I'll have even more. So it's just, again, you know, annual expansion. But the methodology is identical.

CO-CHAIR ROSENZWEIG: And you actually are reporting this on 88 percent, I read in there?

MR. HAMLIN: Yes. It's a high percentage, yes.

CO-CHAIR ROSENZWEIG: Okay.

DR. HWONG: Good. So do we need to do some voting now, or are we like -- should we keep moving?

CO-CHAIR ROSENZWEIG: We've still got 2.B.2, so why don't we vote?

DR. HWONG: Did we do the validity testing?

MS. TURBYVILLE: Typically what we had done earlier was get through the scientific acceptability section and then go

back and vote on the subcriteria. However, I think because 2.A.1 incorporated so many components that we wanted to take a break there and vote. But then you're more than welcome to get through 2.A.2, B.1, B.2 and then go back and vote.

CO-CHAIR ROSENZWEIG: Yes, I think we should -- let's finish --

DR. HWONG: Do you want to keep the momentum?

CO-CHAIR ROSENZWEIG: Yes.

DR. HWONG: Okay.

CO-CHAIR ROSENZWEIG: At least until usability, okay?

DR. HWONG: Okay. That's fine, yes, I think we're almost there then.

So you know, I wanted to say sort of, I think we had this discussion a little earlier, sort of a little bit, you know, in terms of validity testing and I was sort of commenting on how there was a lot of face validity opportunities, you know, back and

forth in terms of modification of this, you know, with their clinician groups.

There wasn't -- you know, like I said, my only concern has simply that, again, using this sort of standardized pricing, how that actually, you know compares with the actual, you know, contractor grade, was it costs, et cetera. I think, you know, would be something of use. But that's all kind of I had, as far as comments go, in terms of validity testing. I didn't know if anybody else had anything else they'd like to contribute in that regard, or even anything from the measure developer in that regard?

(No response.)

DR. HWONG: Okay. So let's go back.

But let's go to 2.B.3 then. So that would be exclusions supported by clinical evidence or otherwise. And we've already talked a little bit about this. Those key categories that I think we mentioned earlier.

And the only thing that when I looked at this, and I said, sure, you know, you look at this and these are expensive categories: cancer, ESRD, et cetera.

But I wondered if there's a way to kind of continue to update that, right, in terms of -- you know, I sort of look at that and I think it could be set in stone and ends up sitting there for ten years, right? But is there something else that's much more of a sort of empiric kind of way to say, hey, you know, when we look at historically and look at patients who have these X, Y, Z conditions, they actually constitute the top, you know, one percent of costs, et cetera? Or this is -- you know, these become, you know -- that are unrelated to these conditions, but these other entities, and we'll reassess that periodically, every five years or every, you know, several years?

So it's just one of those things, when I look at this, and I think it gets kind

of codified, right, as like, okay, great, these are sort of expensive, you know, areas, but doesn't necessarily -- I don't sort of see how that gets sort of readdressed and what sort of empiric criteria was applied to say, you know, these are the conditions, you know. Again, sort of -- you know, start off being sort of unrelated to the condition at hand, right? The diabetes. But you know, are actually significant contributors to cost?

MR. HAMLIN: Yes. These measures undergo effectively almost an annual analysis. The standard prices get updated every three years, and so we do do some analysis as to sort of conditions and costs and sort of reliability and validity of the data, albeit it's not really a deep dive into it. But we do look at these things.

I mean, these conditions were really effectively selected at the get-go. They were shown to be the ones that had the greatest effect. We haven't done updating

since then other than to the codes associated with those conditions. Those get updated, again, every year. You know, perhaps we could take another sort of deeper dive into the exclusion criteria and think about, you know are there other conditions or could one of these conditions maybe be modified to only include stage four ESRD or something like that.

DR. HWONG: Right. And maybe some cancers that maybe are fairly benign and with some treatment are actually -- you know, would not be an actual high, you know, cost outlier, et cetera.

But I just was, you know, for me I thought, it would be nice to hear about some way that you can kind of continually update that and sort of have a little more of an empiric basis to say, you know, why you think these are high cost drivers, right, as opposed to --

MR. HAMLIN: We actually have a

public comment system where we -- sort of every day of the year we actually get comments on codes and conditions from the general public, and their positions or plans or whatever. And they actually do an annual recycle. So they do it again, they get updated every year. We just haven't -- those four have been kind of the, you know, the stone fort, I don't have any other way to --

CO-CHAIR ROSENZWEIG: Which four?

MR. HAMLIN: HIV, organ transplantation, ESRD and -- I did them out of order so I forget --

DR. HWONG: Cancer?

MR. HAMLIN: Active cancer, thank you.

CO-CHAIR ROSENZWEIG: Now, the interesting thing is that two out of those four are clearly related to diabetes. I mean, about 50 percent of kidney transplants have diabetes nationally --

MR. HAMLIN: Yes.

CO-CHAIR ROSENZWEIG: -- and a significant number -- and almost all of pancreas transplants currently have diabetes.

MR. HAMLIN: Yes.

CO-CHAIR ROSENZWEIG: And the other one you mentioned was ESRD.

MR. HAMLIN: ESRD, yes.

CO-CHAIR ROSENZWEIG: About 50 percent of the people on dialysis in the country have diabetes, so -- 60 percent, probably closer to 60 percent. So they're not unrelated. And now there is more and more data showing various cancers being associated with diabetes.

MR. HAMLIN: Well, I think, too -- I mean, again, these apply across all five of our chronic conditions, and that's the way it was designed to have steady methodology. But I think, again, there may be a logic behind, you know, making them required for certain measures where they're less applicable versus diabetes, where, yes, I would agree, ESRD

definitely, and organ transplantation may, at least, in some, you know, maybe in kidney transplantation should be in there whereas others should not be. Yes, I agree with you.

DR. PALESTRANT: This is David Palestrant. How is the age criteria defined for --

DR. HWONG: Sorry, I think we're having a hard time hearing you. Could you get closer to your telephone?

DR. PALESTRANT: Can you hear me now?

DR. HWONG: Yes, better, thank you.

DR. PALESTRANT: The problem I have is with the age criteria. And it applies, like you said, in your other chronic disease categories, but after 75 years of age, all these diseases increase in prevalence, and also this is where most of the costs come, especially for Medicare.

DR. HWONG: Yes, I'm sorry, we're

still having a problem -- I apologize. Maybe try again right now?

DR. PALESTRANT: All right, can you hear me now?

DR. HWONG: Maybe a little slower, too, would be helpful, too.

DR. PALESTRANT: I have an issue with the age criteria. And this is not just applicable to this -- to diabetes, but to your other chronic diseases as well.

After age 75, as you know, all of these diseases increase in prevalence and that's also where the majority of cost is, especially for Medicare, which is one of the areas that you're looking at. So it seems sort of strange to me that 75 is your cut-off.

DR. HWONG: So to summarize, your concern is that 75, you know, may be too low a cut-off in terms of the upper age limit?

DR. PALESTRANT: Right. You know, as people get -- you know, the average life expectancy for a male in this country is

somewhere around 75. So half the patients basically are going to be above that age where they die, and it's also going to be where the highest use of health care resources is within the last three months of life.

So if we don't include those patients, I don't think we're getting a true -- you know, it's consistent, I agree. But I would think we'd want to measure the time when the utilization is going to be most costly and most resource intensive.

MR. HAMLIN: Yes, that's a comment we frequently hear. Again, this is an issue of the criteria being a derivative of our HEDIS product. The original justification for the 18 to 75 under HEDIS effectively was the care side was that management of diabetes over the age of 75, when these measures were developed, became much more complex. And so we didn't want to be attributing much more complex management of people with chronic diseases to levels that were maybe more

appropriate for some of the younger population.

And again, there would be relative resource use measures reflect that, again those HEDIS definitions.

The current HEDIS effectiveness of care diabetes group is looking at age criteria as now diabetics are becoming older and older and not having as many comorbidities or complications, and so therefore they're examining the upper age thresholds. Should that happen on a diabetes HEDIS side, it will definitely be reflected in the RRU side and we'll probably increasingly be incorporating geriatric and the older populations in our measurement strategy.

But for now, again, we're just -- we're in line with the current HEDIS and cost approaches.

DR. HWONG: Yes. Personally I think there are some advantages to be able to kind of line that up with the quality. I

mean, this is the whole concept, right, tagging the quality measure with the actual, you know, resource use measure. So I see sort of the advantage of that.

The interesting thing, and you know, I also agree sort of maybe later on if there are sort of opportunities to do sort of different -- again, some of these categories, break it out, but sort of for specifically, like if you're reporting on the Medicare line of business, you know, doing only the 65 to 75, you know, you could potentially expand that. And especially for, you know, health plan indications or sort of the Medicaid rates or the SSB kind of state sponsored business population. You know, a pediatric, you know, version maybe, you know, of interest, especially, you know, in terms of the growing incidence of diabetes.

So you know, some interesting things that kind of come out of thinking, you know, about the age sort of criteria that --

MR. HAMLIN: Yes, there's actually a pediatricians' group right now working on that. And gain, if those were developed for HEDIS, we would immediately include that as perhaps an age strata in the diabetes, like we do for asthma.

DR. HWONG: Excellent, thank you.

MR. HAMLIN: So they're working hard.

CO-CHAIR ROSENZWEIG: After age 75, a lot of the non-diabetes related medical issues swamp out the diabetes-related medical issues. So you're dealing with end of life care, a lot of very expensive kinds of issues with Alzheimer's disease, a whole variety of conditions that come up that consume a large part of the health care dollar that can be less specifically diabetes related.

MR. HAMLIN: And again, after that age it became too hard to sort of -- to parse out the ones that were non-diabetes related and the ones that were. And you know, we

don't like too much complexity in our measurements, so we want to keep things as clean as possible.

DR. HWONG: Okay. Good, all right. So --

MS. TURBYVILLE: Can I add just one --

DR. HWONG: Yes, go ahead.

MS. TURBYVILLE: Just as a reminder, we did hand out today, as well as emailed earlier, your -- in addition to the lead discussant's comments, your other colleague's discussions. So feel free to look at them and provide your input or whether or not your input is in here as well during the conversation.

I just wanted to encourage everyone to do that if it feels right.

DR. HWONG: Okay. Great, so why don't we move on to -- I guess we're on 2.B.4 then, right?

So evidence-based risk adjustment

strategy. I think we spent actually a fair amount of time talking about the use of the HCC categorization and how that's assessed and sort of ultimately reported out. So I don't know if we need to go over that any further, unless anybody has another comment?

(No response.)

DR. HWONG: Okay. So I'm going to move this along.

Let's go to 2.B.5, all right, which is methods for scoring allowed for identification of statistically significant or meaningful differences, right. And again, I think there was some good thought put into this. You know, there's sort of a standard sample size, you know, that's required of greater than 400 member cases. You know, you're sort of calculating through observed to expected ratios. And then, you know, in essence sort of the larger numbers you have is sort of standard of error that you'll probably have.

So I think you -- as far as when I looked at this, there was a way to actually statistically show, you know, whether someone is an outlier or not, you know, in terms of this. And so there's a very large sort of sample size, you know, requirement there.

CO-CHAIR CURTIS: Can I ask a question? Is there any vulnerability to gaming in this type of measure if the payers are submitting their claims to you? How do you know it's a complete set?

MR. HAMLIN: Every data submission to NCQA has to go through a very rigorous audit process before they can actually submit the data to NCQA. And so that it's a very well mapped and each auditor has to be certified by NCQA before they can even market themselves as an auditor for this data.

So we have fairly high confidence in the data that's being submitted is -- I'm sure there probably are some opportunities for gaming, but it's minimized by the fact that we

have this very detailed audit process for submission of data to NCQA before we even use the results.

That's it. I'm not going to add anything more, Sally. You were a former auditor, so you know.

DR. HWONG: No, no, no, I think that's a very interesting aspect of this measure, that, that's in place, right. Because in terms of the validity or just from the data integrity from these other measure developers or whatever, I don't -- you know, the fact that there is this threat of the audit -- and I know that health plans are always thinking about this, right, but -- every year, right?

But you know, I think that's an interesting way to kind of, you know, make sure about the integrity about the claims that are submitted.

Okay. Yes, any other comments on that area?

(No response.)

DR. HWONG: And I want to try to move us on to -- if not, let's move on to 2.B.6, which I think actually is not applicable. This is the if multiple data sources are specified to demonstrate that, you know, we're producing sort of comparable results. Again, it looks like administrative claims data is the only source of data at this time for this?

MR. HAMLIN: Yes.

DR. HWONG: Okay.

MR. HAMLIN: And anyone using proprietary coding systems has to map to the --

DR. HWONG: Has to translate that?

MR. HAMLIN: -- administrative code. So even for EHRs you have to have a specific code for mapping to the diagnosis codes.

DR. HWONG: All right. So moving on to 2.C, right. Disparities in care. You

know, if they are -- have been identified, you know, measure specs allow for this identification through stratification of results. So I know that when we looked at this in terms of the reporting, there's a way to stratify against age and gender, which is coupled into the HCC categorization.

But -- and I think maybe this was that whole broader topic about how good are resource measures in terms of these units of analysis to kind of get at sort of disparities or very -- you know, disparities of care in terms of like ethnicity, socioeconomic, you know, background, et cetera.

So you know, I think the best you can say for this one is, you know, age and gender. But it may not get at, you know, some of these other important aspects. But I didn't actually see any other measure developers have a better solution to that, either.

MS. TURBYVILLE: Yes, I didn't as

well. And to -- no, in reviewing the measures. No.

DR. HWONG: Good. So I think we've made it down to three. So if I'm not mistaken, we've gone through all the twos?

MS. TURBYVILLE: Are you ready to vote on all these subcriteria from A.2 on.

CO-CHAIR ROSENZWEIG: Okay. So now we're going back to 2.A.2.

DR. HWONG: 2.A.2.

CO-CHAIR ROSENZWEIG: All right. So 2.A.2 is -- all right, so 2.A.2 on the bottom of page 2. All right, the reliability testing, showing that the results were repeatable. And same -- repeatable and also reproducible.

DR. HWONG: I rank that as high. Again, so I decided the fact that have we have multiple years of measurement, 90 percent of the plans stayed within the, you know, one quartile. Most did not actually have, you know, a shift, you know, within quartiles at

all. So I felt like there was good evidence of this.

MS. TURBYVILLE: Okay.

CO-CHAIR ROSENZWEIG: Okay. So we already did 2.B.1, I believe, so we're going to go to 2.B.2, validity testing demonstrates that the data elements are correct and they reflect the cost of care or resources provided adequately, distinguishing higher and low cost resource use.

DR. HWONG: Sure. So I -- again summarizing, I thought this was sort of moderate. There was a lot of face validity testing in terms of, you know, the modifications of this, looking at sort of conditions you think would be more expensive than others. But again, I just sort of have this, again, from standardized pricing, you know, kind of perspective. You know, I wonder if there is some kind of way to get -- be able to capture that a little bit, you know, further or whatever. So I gave it a moderate

rating.

CO-CHAIR ROSENZWEIG: Did our statistician have anything else he wanted to say about it or -- oh, he's busy.

DR. WEINTRAUB: I don't quite understand your criticism.

DR. HWONG: Sure, yes.

DR. WEINTRAUB: I guess your criticism reflects your -- I'm sorry.

I don't understand your criticism, why you're rating this moderate. I seem to be on the higher end of criticism around here. But I'm not sure what else they could realistically do. You have to make some choices about standardized pricing, or something much, much more difficult. And perhaps less generalizable. So I'm not quite -

DR. HWONG: Yes, no, no. I mean, I'm not -- I think there are advantages of using the standardized pricing tables and sort of the technique as being presented. I think

it was much more about the validity testing to say, you know, have you analyzed -- you know, and maybe I'm, you know, this is completely irrelevant. But I was thinking like, you know, it would be interesting to see how this looks if we were to compare this to just actual costs, you know, for a particular plan.

DR. WEINTRAUB: But there is no such thing.

DR. HWONG: Okay.

DR. WEINTRAUB: I mean --

DR. HWONG: Maybe that's where I -- you know --

DR. WEINTRAUB: Unless you're thinking of a large sample, you can do microcosting. Well you know, very, very hard to do.

DR. HWONG: Yes. So that was my only reason that I thought, you know, perhaps there could have been some more discussion or evidence of that, and that's the only reason I ranked it as moderate.

CO-CHAIR ROSENZWEIG: But there are a lot of -- I mean, with respect to coding, there are a lot of errors that are made, that they have to kind of take into account, as a part of what they're doing. I mean, is that --

DR. WEINTRAUB: That's another story entirely. Misclassification coding errors where you're using claims data here, it's -- is you know, it's pretty -- we all know it's pretty dirty stuff.

DR. HWONG: Right. I guess -- and again, I was just sort of thinking from validity, right, like how valid is this assessment, and it gets down to kind of the data elements and the costs, right. And then do the costs truly reflect kind of the resource utilization that's happening in a given system based on their contracted fees and, you know -- you know, I think that was sort of what I was trying to get at.

But again, I'm sure everybody has

the opportunity to vote. So okay, so maybe we can move on on that.

CO-CHAIR ROSENZWEIG: Okay.

DR. HWONG: So go ahead. Okay.

CO-CHAIR ROSENZWEIG: People I guess took your point of view in mind.

(Laughter.)

CO-CHAIR ROSENZWEIG: All right. So let's go to 2.B.3. The issues of exclusions, that they're supported by the clinical evidence, and you went through the exclusions to a certain extent. And that the scoring include computing exclusions so that the effect on the measure is transparent.

DR. HWONG: Right. So I was kind of borderline moderate/high, but I think I've gone to high, given some of the, you know, explanation and reasoning. But again, my whole concept was just to make sure that nothing got sort of codified in stone and why these and, you know, what was making these particular, you know, four entities, you know,

kind of highlighted, and that there was an empiric method.

So you know, given that there is sort of this annual process, whatever, I felt more comfortable with this review process. And so I've, you know, put my ranking as high on this one.

MS. TURBYVILLE: Waiting for one vote.

DR. HWONG: There you go.

CO-CHAIR ROSENZWEIG: Okay. And 2.B.4, which is --

DR. HWONG: Evidence based, yes.

CO-CHAIR ROSENZWEIG: -- evidence based, yes.

DR. HWONG: This is just some strategy.

So I ranked this as high, given our extensive discussion in terms of the HCC categorization and how that would be applied and also used in reporting, stratifying on those -- in those groups so that you can

compare.

MS. TURBYVILLE: Someone's vote didn't register.

CO-CHAIR ROSENZWEIG: Got to press harder.

MS. TURBYVILLE: Yes. There you go.

CO-CHAIR ROSENZWEIG: I'm surprised there aren't cars that are -- you know, haven't had their alarms going off.

(Laughter.)

DR. HWONG: Or unlocking ones as we speak, right?

(Laughter.)

CO-CHAIR ROSENZWEIG: All right, 2.B.6 --

DR. HWONG: I think 2.B.5.

CO-CHAIR ROSENZWEIG: 2.B.5, you know see, I'm just trying to get ahead of myself here.

All right, okay, scoring analysis allowed for identification of statistically

meaningful -- significant and practically clinically meaningful differences in performance.

DR. HWONG: Right. So I mean, I ranked this as high. Again, the unit of analysis on the health plan level, you have over 400 observations. You know, in terms of individual member observations in order to even qualify for the measurement. And you know, again, so given this sort of vast amount of data, I think you're able to sort of calculate, you know, when health plans are sort of statistical outliers or not.

DR. WEINTRAUB: Do you have examples of that in the literature from the developer?

MS. TURBYVILLE: There's a sample report that they submitted, if you want I'll quickly pull that up.

CO-CHAIR ROSENZWEIG: Which one is it?

DR. HWONG: Right, and maybe --

CO-CHAIR ROSENZWEIG: S-A-D-D? No.

DR. HWONG: It's in the -- yes. The measure developer has --

CO-CHAIR ROSENZWEIG: No, that's not it.

DR. HWONG: -- anywhere in terms of the focus of that, I'm not sure which one would the best descriptor. I was sort of more -- sort of summarizing conceptually kind of what I was reading.

MR. HAMLIN: Right, and I think we gave you two in our sample reports. One was the chart of the plan by plan comparison of sort of the total medical ratios, if you will, of everything -- our revenue of one, which is what you're seeing there. That's an individual plan detail.

So this would be one plan. So each of those different service categories that you're looking at -- this is an older report, I just pulled this example from an

older report. So you're looking at --

CO-CHAIR ROSENZWEIG: This is real data, this isn't just a mock-up?

MR. HAMLIN: Exactly. I think it's like 2008 data or something. This is, like I said, older data to identify the plan.

So each plan gets a detailed report, which is national and regional results. The ratios you're seeing there are the total medical ratios, so you can see sort of -- the numbers are so small I can't really see it from back here. But effectively, a one is the normalized mean. And then the, you know, plans for each of those categories that are below one have lower resource use, that are above one have higher resource use.

So the quality composite is the HEDIS measures that we use that are relevant to diabetes. There are ten measures that were used for this one out of basically the entire diabetes composite. That creates the quality composite, and then the total medical is

actually the resource use ratio. And then the pharmacy is, again, the resource use ratio. And then below that are all the subcategory components of the total medical. So you can see the inpatient/outpatient breakdown.

CO-CHAIR ROSENZWEIG: This is one plan solely?

MR. HAMLIN: One plan.

CO-CHAIR ROSENZWEIG: No, no, go further down, there's another --

MR. HAMLIN: So also included in these results are the scatterflies, where you see -- so this is where you see for each plan, the research use map with quality, and the dotted lines are the -- are the mean of one. So that's where you see the variation in the resource use and the quality. And this is what shows up, like, you would see commercial HMO's for diabetes would be what this would -- would probably this would be included.

And the plan, can you see the red -- yes, the red plan would be the one plan

that was selected, when you selected a scatter plot. And like I said, any plan that falls above the 1.5 or below .5 would be Winsorized, and it would show as a little diamond on the edge of the graph where it was Winsorized to. But any plan that was less than .3 or greater than 3 would not show on this graph at all because those would be considered outliers.

CO-CHAIR ROSENZWEIG: And how many of your HEDIS measures did you use for the quality composite?

MR. HAMLIN: The diabetes quality composite, I believe the most is ten. Just here it's nine because we just converted the blood pressure measure to first-year measure status. But it's the greatest number of quality measures, yes. We use the whole diabetes quality composite, so we have the best -- this has the best variation in the quality component because it has so many indicators in it. Asthma and hypertension as one, so there -- you don't see as much in the

quality variation. But even so, you still see broad resource use, interestingly enough.

CO-CHAIR ROSENZWEIG: That's very impressive.

DR. HWONG: Yes. Okay. So I rank that as high for this meaningful -- you know, find meaningful differences and representing that as such.

CO-CHAIR ROSENZWEIG: Okay. And 2.B.6 --

DR. HWONG: I think we said that doesn't apply.

CO-CHAIR ROSENZWEIG: That doesn't apply, right?

DR. HWONG: Yes, multiple data sources is not applicable.

And then I think 2.C -- I don't know what we've been voting on, again -- sort of across again. That's just sort of this age and gender and we're not able to take into other considerations given the sort of administrative claims data limitations.

So I mean, I sort of -- I sort of previously just ranked it as moderate because you can only do age and gender, but that may be the same for all measure developers.

CO-CHAIR ROSENZWEIG: And region, yes.

DR. HWONG: And I guess region, yes.

MS. TURBYVILLE: We'll be sure to capture that rationale in the report, too.

DR. HWONG: Sure.

Okay. All right, so I think we're ready to move on to items three and four, collectively, right?

CO-CHAIR ROSENZWEIG: Correct.

DR. HWONG: Which hopefully we can get through fairly quickly.

So let's see what I have in here. So we start on like 3.A, right? Let's see, measure performance results are reported to public at large and national community reporting. I rank this as high. You know,

these relative resource use measures are reported in "Quality Compass" and "Annual State of Quality Health Care" reports, so they're visible. They have had sort of an audience and, you know, they're -- yes, I mean I think they have high visibility and they are reported to the public in these publications.

I'm sorry, so let me move through the rest of 3. So results are meaningful, understandable. You know again, sort of representation of the results are done sort of in a very clear way.

And then NCQA mentioned in the documentation that they have a series of webinars, resource documents, they frequently go back for stakeholder feedback, et cetera, to inform -- you know, inform how this information is ultimately presented. So I thought that -- I ranked that as high as well.

Transparency, I said was high as well. They're very clear regarding their specs and process. I appreciated the detail

that they went into, you know, both in the standardized pricing tables as well as the HCC, you know, methodology. So I thought -- you know, so what was represented here would allow for, you know, physicians and health plans or large groups to kind of understand how the score was ultimately generated. So I ranked that as high.

I think 3.D is not applicable right now because that's harmonization.

Again, part of these advantages I think is simply because they have been in use for like the last four years now. So they're visible, you know, there's been opportunities to have that kind of feedback to make those results more understandable.

Okay, shall we vote?

CO-CHAIR ROSENZWEIG: Yes, I think -- okay, so let's go to 3.A. We're talking about current use and --

DR. HWONG: That these are reported in public at large, right?

CO-CHAIR ROSENZWEIG: Yes.

DR. HWONG: And I ranked this high, you know, for the publications that these are reported in.

Great.

CO-CHAIR ROSENZWEIG: Finished?

MS. TURBYVILLE: Yes.

CO-CHAIR ROSENZWEIG: And then the -- are the results meaningful, understandable, useful for the intended audience for both the public reporting and performing quality improvement.

DR. HWONG: I ranked this high. So you know, again, webinars, resource documents, stakeholder, you know, feedback.

MS. TURBYVILLE: Seven high, one moderate, right?

CO-CHAIR ROSENZWEIG: Okay. So the third one is that the data results detail are maintained so that the measure and the logic for defined unit of measure can be decomposed to facilitate transparency. So

it's -- and it can also be taken down to its specific components.

DR. HWONG: Yes. So yes, I ranked that as high, you know, given some of the detail of the specs. And again, I try and view it from, you know, if you're a practicing physician or a health plan, you know, within a group, you know, how do you get to this score? And I think it is -- well, there is a lot of work involved, I think it is fairly transparent.

DR. WEINTRAUB: Well, an individual physician interpreting this would do it within the context of a plan, since they're already proposing that --

DR. HWONG: Yes, right. Within the -- exactly. This is at a plan -- yes, right, at a plan level.

CO-CHAIR ROSENZWEIG: And we also saw a copy of the reports.

DR. HWONG: Yes. Great.

CO-CHAIR ROSENZWEIG: And then --

DR. HWONG: Okay, so then --

CO-CHAIR ROSENZWEIG: You said we did not need to --

DR. HWONG: 3.D is not applicable, harmonization, at this time. So I think we can get through 4 very quickly; 3 and 4 are sort of -- kind of run through that.

So feasibility -- 4.A is feasibility, is the data generated as a byproduct of the care process? And indeed, it is. This is administrative claims and billing. So that I ranked as high.

Is this an electronic source? It is, because it is administrative claims. So that is high as well.

Susceptibility to inaccuracies, errors are minimized. And this is where I sort of noted that the fact that NCQA conducts these audits is well known, people have to prep for that. That I think in terms of like, you know, whether its willful or just neglectful kind of errors in terms of the

data, I think there's just a lot of apparatuses in place at health plans to actually, you know, make sure about data integrity and what's sort of submitted.

So I ranked that as high, simply because there is this auditing arm.

DR. WEINTRAUB: Well, you know, you do have a problem with quality of -- with it being administrative claims data. So I mean, it isn't --

DR. HWONG: The susceptibility is to inaccuracies.

DR. WEINTRAUB: -- I mean, forever -- you know, I mean, you're forever limited by the inaccuracy of that. So what we say about that, I don't know.

DR. HWONG: I would say --

DR. WEINTRAUB: I would move you to moderate rather than high.

DR. HWONG: Right. I would say given the other -- you know, again, all the other measures that we're considering, they

were all administrative claims data so they would all be subject to the similar sort of that, you know, versus electronic medical record, et cetera. So you know --

DR. WEINTRAUB: Or clinical databases.

DR. HWONG: Right. Oh, yes, right. Yes. Registries or something, right?

Yes, I mean, I still ranked it high. Again, I thought it's great, you've got this auditing arm. Again, it's well known among health plans to sort of anticipate that there are resources sort of -- you know, devoted to that.

So are we voting? Oh, I have to go through -- sorry.

And then the one more, just one more. Okay.

So 4.D, data collection measurement strategy can be implemented as demonstrated. And you know, I think the proof is in the fact that they are currently being

implemented and have been, you know, as such for like the last, you know, four years.

So all high for me, in terms of the four subcategories for this -- for number 4.

CO-CHAIR ROSENZWEIG: Okay. So let's vote.

So 4.A is required data is routinely generated and used during care delivery.

Okay, 4.B, electronic sources.

Okay, 4.C, susceptibility to inaccuracies, errors, unintended consequences.

Okay, and then implementability - if that's a real word.

Feasibility, that's the proper word.

All right, thank you very much.

DR. HWONG: Thank you.

MS. TURBYVILLE: Should we take a break and then come back in 10, 15 minutes? I don't know, are breaks usually 15 minutes?

In 15 minutes, and then we'll pick up another measure.

CO-CHAIR ROSENZWEIG: Which one?

MS. TURBYVILLE: I'm trying to remember what we agreed to before lunch.

CO-CHAIR CURTIS: We've changed since then, so I think we should probably just go back to the -- we'll discuss it.

(Whereupon, the above-entitled matter went off the record at 3:45 p.m. and resumed at 4:02 p.m.)

CO-CHAIR CURTIS: So we're going to get started. We're going to mix it up one more time, sorry Ashlie. We've lost the team.

We're going to go ahead and move to the other ABMS-REF measure, which is the -- so 1573, episode of care for management of CAD post-revascularization. And Bill Weintraub's going to lead and I'm the co-discussant.

DR. WEINTRAUB: Okay. This one is episode of care, management of coronary artery post-revascularization. This also comes from

the American Board of Medical Specialties Research Education Foundation. So this is very similar to the measure that Jeptha presented in such detail this morning.

So I'm going to repeat it in exactly the same level of detail and we'll be here until 9:00 o'clock tonight.

(Laughter.)

DR. WEINTRAUB: No, actually I think because it's so similar, they go about things so similarly that we do not need nearly the level of discussion and detail.

Once difference is, unless I'm missing something -- I'm probably missing a lot -- is this is post-revascularization, after the discharge rather than the 30 days to a year. I think that's really okay. They have inpatient and ambulatory facilities under resource use categories. Looks pretty reasonable.

This is also, by the way, entirely based on claims data and it's constructed in,

as far as I can tell, exactly the same way as the AMI measure. So to identify patients post-revascularization, count their resources. And then the costing approach is going to be also exactly the same.

So the testing and reliability are not yet done since this is -- they just made this up. And so I think it is actually a very important issue, important measure of considerable societal interest. We spend a lot on revascularization. We know that there is some variation in care, after people go home after revascularization. For instance there's variation in use of imaging, it's all very uncertain how to go about that. We know a fair amount about disparities of people coming into revascularization, we don't really have much in the way of descriptive data about variations in disparities after.

We know about regional variations in use of revascularization, I don't know much about -- I don't know if there's much

literature on variation use after. But we know there is at least uncertainty as to what things we should be doing with people post-revascularization.

So that's sort of where this is. So I guess we can go through all the first-level measures. So I think that this measure is a measure that makes sense and I think it will be of some value once they're done.

DR. MARWICK: Do they include both surgical and post-surgical?

DR. WEINTRAUB: Yes. So it's essentially two measures. They're stratified by that, but they're so different that they're really quite different.

CO-CHAIR CURTIS: So I just want to make sure so people -- it is different in that this is not -- it's more of a stable CAD population. They've excluded patients with an AMI in the preceding year, I believe -- I've got to look that up again.

DR. WEINTRAUB: No, it's true.

CO-CHAIR CURTIS: And so they identified that, in this case, the index is not an AMI admission but rather the performance of some revascularization procedure, CABG or PCI. And then the follow-up period is -- you know, starts, I guess, at discharge, although it gets a little unclear to me.

DR. WEINTRAUB: That's a little weird, but that's certainly my take.

And then the other thing -- certainly the category patients with this would be described as stable ischemic heart disease, is what's become a popular term these days.

MS. CLARK: And CABG and PCI are combined together?

DR. WEINTRAUB: No.

MS. CLARK: There's two separate --

DR. WEINTRAUB: Essentially it's two separate -- doing essentially the same

thing.

CO-CHAIR CURTIS: I thought they were specifying it as one, but including both populations. I didn't think they got so they were going to stratify it.

DR. WEINTRAUB: Yes? Well, all right, it may be just my take.

DR. REEDER: What's the level of analysis?

DR. WEINTRAUB: What do you mean, what's the level? Do you mean is it --

CO-CHAIR CURTIS: Provider?

DR. WEINTRAUB: Oh, I see, provider. They are unclear.

DR. REEDER: It says per episode of care.

DR. WEINTRAUB: They say anywhere from the individual practitioner to the health plan level. They're unclear about that.

DR. REEDER: Okay.

DR. WEINTRAUB: Similar to this morning's discussion with the AMI measure.

MS. TURBYVILLE: I just want to say that they're clear that they believe this measure could be used at a individual provider level. It could be used rolled up to a health plan as specified. So what is up to you is, as specified, can it be implemented and does it make sense, et cetera. So yes, right.

DR. WEINTRAUB: Right. So this is my discussion this morning about, as Jeptha was presenting it, it's the same. That they say they can do it, yes, from the level of the provider through the health plan. And I think it seems unlikely to me.

DR. REEDER: Thank you.

DR. WEINTRAUB: So it seems to me --

CO-CHAIR CURTIS: I think we took the overview, and take a step back to go to importance, which I think you --

DR. WEINTRAUB: So do we want to discuss this more? Or both?

CO-CHAIR CURTIS: It's probably --

I don't know. I don't know if people are clear on how they specified the importance of it. But I think it does overlap with the previously discussed measure, in that this is a CAD high-impact area. So that's easy. The disparities, yes. So to stick to your bottom line, I think we can use the same criteria and the same discussion from this morning to vote on importance. But I want to make sure it's clear in everybody's head, exactly where we're at.

DR. WEINTRAUB: And I rated the importance of this as high.

CO-CHAIR ROSENZWEIG: Timeframe for revascularization?

DR. WEINTRAUB: One year.

CO-CHAIR ROSENZWEIG: Thirty days to one year or just --

DR. WEINTRAUB: Not entirely clear, but it looks like it's discharge to one year.

CO-CHAIR ROSENZWEIG: Discharge to

one year?

DR. WEINTRAUB: Yes.

CO-CHAIR ROSENZWEIG: Okay.

DR. WEINTRAUB: That's not unreasonable.

MS. TURBYVILLE: I think Tom has a question, too.

DR. MARWICK: Yes. And just in terms of their description of the impact, they don't actually talk about a post revas group, they talk about chronic CAD.

DR. WEINTRAUB: Well, these are all patients that have had revascularization.

DR. MARWICK: No, no, I agree. But if -- it goes back to our discussion this morning as to what we're scoring. Are we scoring what we consider to be the importance of the topic, or how they are portraying the topic? And my understanding is that we're scoring how they're portraying the importance of the topic. And their description of the topic is actually talking about chronic CAD,

not about the numbers of people that get revascularized and how they're followed up and whether they're managed inappropriately, or whatever.

CO-CHAIR ROSENZWEIG: That's a good point.

DR. REEDER: It says on page 5 this measure can be used to identify, and if necessary address unwarranted variability in the resources used to treat CAD patients post-revascularization annually. So it is the treatment of patients who have CAD, but who also have been revascularized from discharge up to 12 months.

DR. WEINTRAUB: Well, I think our -- I have a feeling we're dancing on the head of a pin here. These are patients who, with chronic stable ischemic heart disease - really I think they did pretty well, actually, which is okay. These are patients with stable ischemic heart disease who are revascularized. And then from that point forward, we are

looking at resource use.

DR. MARWICK: I understand, we're completely on the same page about what this is about.

My observation is about, if we're giving the imprimatur to how this is put together, then the background information that they've provided is not about a post-revascularization population, it's about chronic CAD. So if we're scoring the importance of it based on how they're spinning the importance here, they haven't actually done a very good job.

CO-CHAIR CURTIS: But I would argue, I agree, but I would argue that this is another one of these semi-paired measures where there is a CAD without revascularization or some other cohort that's similar, I believe, that we're going to review tomorrow. Am I correct in that? I believe so.

MS. TURBYVILLE: 1572, I believe is --

CO-CHAIR CURTIS: Right. So I think they're probably just using a generic discussion for the overall stable CAD population, and not specifying. They do cite one paper that focuses on differences in utilization post-VCI.

DR. STROUPE: This is the MS measurement developer, I'm Kevin Stroupe.

That's exactly right, there are two separate measures. One regarding CHF which -- chronic CAD, which was looking at a stable chronic management period for a year period of time, and then there was also the post-revascularization measurement period, which is also trying to look at -- that would be a similar type of population, but following post-discharge for revascularization.

DR. WEINTRAUB: All right. So let me just ask you a simpler question. The period here is from discharge to one year, is that correct?

DR. STROUPE: That's correct.

CO-CHAIR CURTIS: All right, so I'm ready to vote on importance, if everybody else is, before we degenerate.

So getting back onto the specific criteria which I now -- so in terms of the importance, 1.A, the impact, grab your keychains. I think we can agree that this is a high-impact area or disease.

Go ahead and vote.

And then for 1.B, is there a performance gap demonstrated in the data? Bill, how did you vote that for preliminary review?

DR. WEINTRAUB: I voted it high, but I was probably being generous.

CO-CHAIR CURTIS: Okay. So it's --

DR. WEINTRAUB: But I think it probably is, but I'm not sure they've demonstrated it perfectly. I'm not sure the literature would support it. But I think there is that.

All right, we're on the same page here.

CO-CHAIR CURTIS: Let's go ahead and vote on that.

MS. TURBYVILLE: One vote hasn't come through yet.

CO-CHAIR ROSENZWEIG: Do you have any sense as to how much of the results are affected by the quality of care after the revascularization or the actual quality of the revascularization procedure?

DR. WEINTRAUB: Gosh, that's a great question for research. If you can come up with some good ways of studying that, that would be good.

CO-CHAIR ROSENZWEIG: I don't want to get all --

CO-CHAIR CURTIS: I would assume it's much more about the system of health care in which the care is delivered as opposed to the quality of the PCI itself.

DR. WEINTRAUB: You know, resource

utilization is going to be -- you know, it will be driven by -- but I think that selection of patients for vascularization is very difficult. But the quality of revascularization today has almost been commoditized. It's not that -- I don't think it's that variable.

CO-CHAIR CURTIS: So one high, six moderate and one insufficient.

And moving on to 1.C, the intent of the measure.

DR. WEINTRAUB: So the purpose, objective of the resource use measure -- I'll get it right yet, I promise.

The purpose of the resource use measure, including its components, and the construct for resource use costs are clearly described.

It's not perfect, but I think it's high.

(Laughter.)

DR. WEINTRAUB: Am I missing

something?

CO-CHAIR CURTIS: He was trying to vote with his BlackBerry.

(Laughter.)

CO-CHAIR CURTIS: Let's go ahead and vote on --

DR. WEINTRAUB: What? I came to the wrong meeting.

MS. TURBYVILLE: I do love you all. You're very different than the surgeons last week. They just sat there. You all put it down and then pick it back up. The surgeons were like aiming at it.

DR. WEINTRAUB: Maybe they thought it was a scalpel.

CO-CHAIR CURTIS: And finally, the resource use category, whether or not this is -- the appropriate categories are being assessed and consistent with the measure intent, to recap it.

So it's our usual issue. Bill, do you have a suggestion on this?

DR. WEINTRAUB: Yes, I voted moderate. I don't think they explained this as well as they might have. But you can -- I don't know, it's somewhere between moderate and high.

DR. WEINTRAUB: Okay. So we're ready to plunge into the measure. Jeptha took us through in two and a half hours, I'm going to take us through in 15 minutes.

CO-CHAIR CURTIS: You're a better man.

DR. WEINTRAUB: Probably not.

Okay. So their general approach, you've heard for the using claims data to look at resource use in this target population. So as we already heard -- so the patients have to be enrolled for 24 months, similar to what we heard this morning. Describe the data cleaning steps, most of it's fairly straightforward.

CO-CHAIR ROSENZWEIG: Bill if you can give us an idea of where you are on the

PDF?

DR. WEINTRAUB: Okay, sorry. I'm on page 9, looking at the PDF data exclusion category criteria.

CO-CHAIR CURTIS: Let me make a suggestion. Perhaps aside from the population definition, if the measure developer could point out how this methodology is different, if at all, from the previously discussed measure on utilization post-AMI? Because as far as I can tell, as Bill's alluded to, data cleaning, data specifications, cost methodology, everything was identical.

And so in the absence of a significant difference, as I think Bill is suggesting, we can really just use the -- unless there's been a change of opinion, use the same feedback to the measure developer.

So let me open that up to the measure developer.

DR. WEINTRAUB: Good point.

DR. STROUPE: The data cleaning

and that, this was similar to the approach before. The difference is just around the particular fact that, for this measure, there again was the same requirement for the continuous coverage and for this period, though, there were -- for this measure, the exclusions that there had to be, there couldn't have been an AMI between 14 and 365 days, as far as specific criteria.

But then there were the other or a prior revascularization, and then the other sort of standard excluding criteria, the end-state renal disease and cancer and so forth. So basically set for some particular issues around the procedure, prior revascularization or the clinical conditions.

And what was looked at as far as revascularization being the triggering event, otherwise the data source that was used for that, the measurement testing and the other data cleaning and so forth were similar to the other measures that have been described.

DR. WEINTRAUB: All right. So once again, they do not recommend imputing for missing data, that's 6.4 on page 9.

Help me, what am I supposed to do?

CO-CHAIR CURTIS: So I think if the only difference is really the inclusion population, I think there are some areas that are worth reviewing on that, and specifically that one that you mentioned. The exclusion of patients with an MI 14 to 365 days before the index revascularization. So I'm not sure where that is in the --

DR. WEINTRAUB: Yes, how did you pick that? What's your rationale? Why 14 days?

DR. STROUPE: The --

CO-CHAIR CURTIS: This is page 10 of the PDF.

DR. STROUPE: The measurement was -- these criteria were worked through in conjunction with a clinical advisory group. And the rationale regarding that was that an

AMI -- if we excluded AMIs closer to the triggering event, that the -- there might be some AMI that might be associated with the revascularization that we were wanting to capture to be the triggering event of the episode.

DR. WEINTRAUB: So I'm sorry, I'm not following. Are patients included if they have a revascularization within 14 days after an MI?

DR. STROUPE: Yes.

DR. WEINTRAUB: All right, so this is not -- so what I said was wrong, it's not about stable ischemic heart disease, it's a mix?

DR. STROUPE: Well, in that there's that time window when the AMI could have occurred, then leading to the triggering revascularization event.

DR. WEINTRAUB: Yes. So then it's a mix of patients with stable ischemic heart disease and patients with -- who have had a

recent acute MI?

DR. STROUPE: Recent in the -- at least within a two-week period. But prior to that two-week period, they would be excluded.

DR. WEINTRAUB: Right.

DR. MARWICK: Could I seek some clarification as to whether the identifier is the process of getting the revascularization? Because on page 10 in the clinical framework, the first step is to identify patients, episode inclusion criteria of one ambulatory visit for CAD-related care. That sounds like a chronic CAD descriptor rather than a post-revascularization descriptor.

DR. STROUPE: For the patient and the chronic -- in the revascularization measure, the triggering event would be the existence of a -- would be the existence of the revascularization event. That's -- that was the indication that was -- we were looking for to trigger the revascularization measure.

DR. MARWICK: Yes, I get that.

It's just in the framework it doesn't seem to read like that. Or maybe I've misunderstood it. But it seems to -- in the framework, it seems to read like chronic CAD.

CO-CHAIR CURTIS: I think the top level is patients who had a revascularization, either type technique, and in the subsequent 12 months they had to have at least one CAD-related resource use to be included in the measure, is the way I interpreted that. So that if for some reason unknown to man, they went back to France or wherever they were from, and they were never heard from again, that they wouldn't be included in this measure. But that might be an assumption on my part?

DR. WEINTRAUB: Is that correct? They have to have some evidence of resource use in the year after their revascularization?

DR. STROUPE: That right there, there does have to be -- in order to so that we'll be able to capture there, they have to

have some use, right. That's correct.

DR. MARWICK: That's a source of bias, isn't it? Because if your question is about the costs of management in the year after revascularization, then you don't want to limit it to people that have actually -- to only people that have incurred some cost. I mean, there might be some people that haven't incurred any cost.

DR. WEINTRAUB: That would be pretty unusual, to have no resource use at all, unless you're dead or have left the planet, have no resource use at all would be most unusual.

DR. STROUPE: There would be a concern with having zero costs that there might be some data missing issues for those particular --

DR. WEINTRAUB: I agree, and I'm not that troubled by it. I mean, you don't -- they're not asking for very much, they just want to know you haven't disappeared from the

planet.

So all right. So these are patients 18 years and older. The one this morning I think excluded people over the age of 85. But this one does not. And here we see acute myocardial infarction 14 days to 365 days as exclusion, and I missed that.

And then they identify patients by, what is it, CPT codes?

CO-CHAIR CURTIS: Sorry, I just want to sort of build on that. I don't understand that exclusion of sub-acute infarcts or within the one year prior to or not the 14 days prior to. It just doesn't make sense to me. Either this is a revascularization population or it's not. Right now you kind of have the highest-risk patients, the MIs, the acute MIs, you've got the lower risk ones without any MI, but you're missing that middle of the sandwich, which is the remote MI.

DR. WEINTRAUB: Yes, I agree.

That's peculiar. Are they going to make a case for no prior MIs, it's purely stable ischemic heart disease, or you could make a case for including anyone with a prior MI. But then of course, someone with really remote MI is included. So it's peculiar, the 14 days to 365 days.

DR. STROUPE: That two-week window is that, again, with discussion from our clinical advisory panel, there was concern that by eliminating patients with that diagnosis within that two-week period, that you might be excluding a larger number of patients than we would want, patients where the MI then was related to the revascularization just prior to the trigger.

DR. WEINTRAUB: Yes. So I understand. The real question is, why exclude the patients 14 to 365?

DR. STROUPE: Oh, why not exclude the patients 14 to --

DR. WEINTRAUB: Why exclude them?

DR. STROUPE: Oh, why exclude them. The -- from the 14 to the 365, that was to -- so that the population would be, again, one -- trying to get a population that would be more in a stabler or more management phase of the condition.

DR. WEINTRAUB: Then you would want to exclude the zero to 14. I mean, you sort of --

CO-CHAIR CURTIS: You can't have it both ways, I think is what you're getting at.

DR. WEINTRAUB: You can't have it both ways. This is one failure here.

CO-CHAIR CURTIS: So I think, you know, we've heard your response. If in the course of matters you want to respond to this more, this would be part of the feedback that we provide to you, but obviously there's a difference of opinion with the members of the TAP.

DR. WEINTRAUB: Fair enough, why

don't we move on.

CO-CHAIR ROSENZWEIG: Are you excluding patients who have had a revascularization in that period of time? A prior revascularization?

DR. STROUPE: We -- yes, a prior revascularization. A revascularization is the -- that's that the triggering event is then post-discharge on one year of the patients that we followed in this measure.

CO-CHAIR ROSENZWEIG: I don't understand.

DR. WEINTRAUB: So consider the timeframe. Someone is within the timeframe of the measure, they have a revascularization. They're not going to enter again for -- at least for the next year. But suppose it's not within the timeframe of the measure. Can someone enter twice? Can someone be in the measure in year one, have a revascularization, year two they come back and have another revascularization? Can they enter the measure

again?

DR. STROUPE: Within the close revascularization measurement, if someone would have the -- a revascularization within that year -- within that one-year period, that the multiple revascularizations were -- would be allowed in the measure.

DR. WEINTRAUB: That's not the question. So that would be additional resource use for the year, the next year. So year one's over, and now they come back in year two. And three months into year two, they have another revascularization. So in year one they have a PCI and 15 months later, when their first year is over, they have CABG, and they enter again as a new revascularization, would then be followed up for yet another year?

DR. STROUPE: Oh, would that be -- okay.

CO-CHAIR CURTIS: Let me just skip to the chase. The answer is yes, because they

are screening -- you are screening for other revascularization procedures in the preceding 12 months, and that's one of your exclusion criteria. So I think that's pretty clear, that you can enter in a subsequent month more than 12 months set apart.

DR. WEINTRAUB: So that's actually very similar to the MI measurement this morning.

DR. STROUPE: That would be correct. Within that one-year period you can have the subsequent revascularization, but in the following time period, then if you were starting to measure the first one from that triggering event, and looking prior -- for a prior one-year period, then as noted, that revascularization in year one would be picked up as part of -- during the screening process.

DR. WEINTRAUB: But would that kick them out of year two?

DR. STROUPE: Could you repeat that, please?

DR. WEINTRAUB: So if it's not within 12 months but it was in year one, it still could enter in year two?

DR. STROUPE: Within the second year, there's a triggering event in year two and you look back 12 months and there hasn't been a triggering revascularization from year one was more than 12 months before the revascularization in year two --

DR. WEINTRAUB: Okay.

DR. STROUPE: -- then that could be picked up again.

DR. WEINTRAUB: Okay. Everyone got it? Good.

All right. So I'm on page 11. So they identify the patients, and then they look for these exclusion criteria in detail. Then they identify the measure population in step four.

CO-CHAIR ROSENZWEIG: Why would insertion of a cervical dilator be considered?

DR. HWONG: Is that categorized

like under a pregnancy-related --

CO-CHAIR ROSENZWEIG: It's an exclusion criteria.

DR. REEDER: There are a lot of OB/GYN procedures that are in here, and I would ask the same thing.

CO-CHAIR ROSENZWEIG: But someone has a Pap smear, I mean, are -- or a -- why would that exclude them? Or someone has a DNC? If someone has a DNC, that shouldn't exclude them from -- am I mistaken?

DR. HWONG: I got the sense that the exclusion here, at least for all these codes here, was trying to be -- you know, whether or not it's successful this way, but trying to be related to pregnancy. Because pregnancy, in terms of resource utilization, is you know -- can contribute to costs. So I don't know if it's --

CO-CHAIR ROSENZWEIG: Oh, that's all pregnancy related?

DR. HWONG: I don't know, like

we'd have to have a code, or someone who's very adept at these codes to kind of say, like, how often these codes ever associated with pregnancy. I'm assuming --

CO-CHAIR ROSENZWEIG: You're assuming, okay.

DR. HWONG: You know, I would assume that, you know, we could kind of look into that. But if the overall concept here is to exclude, you know, patients who have been pregnant at any point, you know, during this measurement period, then maybe -- I get a sense that they're trying to capture that.

CO-CHAIR ROSENZWEIG: So DNC --

DR. HWONG: Your question is a good one.

CO-CHAIR ROSENZWEIG: So a DNC not related to pregnancy shouldn't exclude you?

DR. HWONG: Right. So we'd have -- so maybe the comment to go back and just to make sure that these are very specific about pregnancy, right?

DR. REEDER: I think it has the V code, the overall general V code for pregnancy in here. So they are trying to get everything.

CO-CHAIR ROSENZWEIG: Okay, all right. Sorry.

DR. WEINTRAUB: So the exclusion just discussed is very nicely highlighted. The rest is very nicely highlighted here, cancer, end-stage renal disease, organ transplant, HIV/AIDS and things related to pregnancy, largely. It's the kind of things we've actually seen before.

They identify their population -- I'm on page 11. And then they look at these various events. And again, this is the kind of framework we saw this morning. Rather than trying to capture everything, we're looking at specific areas of resource use.

CO-CHAIR CURTIS: And it looks like it's completely identical to that used in the earlier measure.

DR. WEINTRAUB: So we have to worry that Mary Ann found coding errors this morning, and they would have to go through this again and make sure, if they're going to do it this way, that they don't have similar errors.

So now I'm up to page 12. They're taking out patients less than 18, they justify that.

Let's talk about their exclusions again, this is really repetitious of what they have above, and they justify their specific codes on page 13.

DR. HWONG: So one thing I was -- on page 13, there is this mention about "to be included in the measure, an individual must have an inpatient admission for heart failure or congestive heart failure."

Oh, I'm sorry, this is on page -- I moved ahead a little bit -- on 12, sliding between 11 and 12 here.

So I just wanted to -- I'm trying

to recall, but this sort of stratification between you have heart failure, you don't have heart failure, was that in the ABMS version or was that in the --

CO-CHAIR CURTIS: We should ask the measure developer for help on this.

DR. WEINTRAUB: Yes.

CO-CHAIR CURTIS: But it's not specified elsewhere, and I assume that that was a typo or an oversight --

DR. HWONG: Yes.

CO-CHAIR CURTIS: -- where they were trying to put seven applications in in a very short time.

DR. HWONG: Sure, yes. So yes, if that's something we can sort of disregard, that's good to know, right?

DR. WEINTRAUB: So let's ask the developer?

DR. STROUPE: That was probably a typo.

DR. HWONG: Okay, good. So

there's no -- nothing about having to have heart failure to be included in the measure?

DR. STROUPE: No.

DR. WEINTRAUB: That would make no sense.

DR. HWONG: I just wanted to make sure. Okay.

DR. WEINTRAUB: So they go through a lot of their logic on page 13 and into 14. They do not -- they're not specifying severity levels.

CO-CHAIR CURTIS: Which I think is in contrast to the other measure, where they stratified by heart failure.

DR. WEINTRAUB: They said, we attempt to create a relatively homogeneous population through our inclusion/exclusion criteria. Well, lots of luck. The variability of patients undergoing a revascularization is extreme, and right now they've got both patients who are recently post-acute MI and patients with stable

ischemic heart disease.

MS. TURBYVILLE: It seemed like they would want to look at stratification based on diabetes status, right? Those are usually --

DR. WEINTRAUB: So you know, I think that, do you want to stratify on it or do you want to be able to examine it as a subgroup? I mean, they really -- I would stratify on type of revascularization, because they're so very different, percutaneous and CABG.

For other things, diabetes, non-diabetes, heart failure, no heart failure, severity of coronary disease, age and gender, those can really be examined in subgroups rather than stratify. That's how I would handle it.

The idea that they're a homogeneous population doesn't make sense.

DR. MARWICK: You would also want to stratify on acute and chronic.

DR. WEINTRAUB: Yes, if they're going to do it that way. Revascularization setting of a recent acute myocardial infarction and stable ischemic heart disease are pretty different.

That being said, I've just gone through this recently about stratifying versus all within one analysis for ischemic -- first, stable versus acute and there's no perfect answer to it.

(Off mic comment.)

DR. WEINTRAUB: No, that was -- it's not in this.

MS. TURBYVILLE: Typo.

DR. WEINTRAUB: That was a typo. Actually, they said it was an inclusion criteria, but that was a typo.

CO-CHAIR ROSENZWEIG: So it's an exclusion?

DR. WEINTRAUB: No.

CO-CHAIR CURTIS: It's just --

CO-CHAIR ROSENZWEIG: So shouldn't

there be some stratification based upon CHF?

DR. WEINTRAUB: So do you want to stratify all these different things, or are they really co-variants or subgroups? I think to stratify on all of the different things doesn't make sense. I mean, stratification means you're only analyzing within the group. So I would say diabetes, heart failure, severity of disease, left ventricular function, age and gender are subgroups without stratifying variables. But I think that logically, type of revascularization is.

And you could argue acute versus non-acute.

CO-CHAIR CURTIS: And so just let me ask the measure developer that question. Is since you are including two very types of procedures, and certainly resource use would -- in the following year would be expected based on clinical trials, to vary by that procedure, did you consider again stratifying your subgroup reporting by procedure? I

didn't see that in the application, but maybe I missed it.

DR. STROUPE: Yes, the stratification that was -- had been so far was stratifying regarding whether the patient had a subsequent or only one revascularization. But that's certainly another group.

DR. WEINTRAUB: I don't follow that. That doesn't quite make sense.

DR. STROUPE: The stratification that was -- that is originally proposed in the measure is to stratify individuals based on following the trigger event, whether they had no subsequent revascularization or whether they did have a subsequent --

DR. WEINTRAUB: That's an outcome, that's not a stratification variable.

CO-CHAIR ROSENZWEIG: Maybe he's thinking prior --

DR. WEINTRAUB: Maybe.

CO-CHAIR CURTIS: But it is -- as currently specified, you guys have

demonstrated your intent to stratify by number of revascularization procedures in the following 365 days?

DR. STROUPE: Right.

DR. WEINTRAUB: Where does it say that?

CO-CHAIR ROSENZWEIG: They say they're not going to stratify --

CO-CHAIR CURTIS: No, they -- so bottom of page 20 -- sorry, bottom of page 20, section 10.2, the patients included in the revascularization measure will be stratified by whether patients did or did not have multiple revascularizations during the 12-month measure period.

DR. WEINTRAUB: That doesn't make any sense.

CO-CHAIR CURTIS: Which I flagged myself as being nonsensical.

DR. WEINTRAUB: That's sufficiently nonsensical.

CO-CHAIR CURTIS: So do you want

to provide the rationale for that? Because I think in most cases, that would be in your outcome of resource use, would be subsequent revascularization.

DR. STROUPE: That there are patients requiring a multiple revascularization, those might require the higher resource use patient that could be looked at separately from the ones that didn't have a subsequent revascularization. For keeping a more homogeneous population of individuals for the analysis.

DR. WEINTRAUB: Well, the problem is that you cannot -- if you're doing that, if you stratify that way, then you can't use it as an outcomes measure. So it doesn't -- it really doesn't make sense. Once you stratified, you can't go back and say, well, now I want to unstratify. I mean, I just don't think that that makes any sense. The stratification variable is essentially always a variable that you have in hand in the

beginning.

DR. MARWICK: I wonder for the developer, if you could get a -- we could get at your intent by stipulating the difference between the people that have had a single and multiple previous revascularizations, you could say --

DR. WEINTRAUB: That's not what it says.

DR. MARWICK: No, I know it's not what it says. But I wonder if that's the intent of the developer, in terms of people that have had multiple previous episodes are more likely to cost more?

DR. STROUPE: Well, with the prior episodes, that would be the case. That they would have been excluded from the measure. Trying again to create a sample that would be --

DR. WEINTRAUB: Within the 365 days?

DR. STROUPE: -- it would be more

homogeneous. This then, just looking at the resource use pattern following the discharge, that those individuals with the subsequent revascularization might have other outcome -- other costs that would -- higher cost profile. And just to keep the groups that would be more similar for --

DR. WEINTRAUB: I think we need to cut this off here or we'll be on it all day. I think this is one of the feedback points we have to give them, that we think there's problems with the construction logic.

MS. CLARK: I wonder about multi-vessel disease, whether that might be an appropriate way to stratify?

CO-CHAIR CURTIS: That is what this is really trying to capture, but it's unacceptable.

DR. WEINTRAUB: Well, it doesn't -- no, it doesn't get at that.

MS. CLARK: No, but they could.

DR. WEINTRAUB: They could, but I

wouldn't advise it. Again, I think that that's a co-variant, or should be looked at as a subgroup.

MS. CLARK: Well, in terms of --

DR. WEINTRAUB: Stratification means that you think that these are really different things and you want to analyze them separately. Otherwise, there's little reason to stratify. So I think you want to look at resource use after PCI and after CABG, you want to look at those separately.

Now you may want to compare across, that's another story. But I don't think that this is not a comparative effectiveness study of different revascularization strategies.

MS. TURBYVILLE: Just a question. Is that the only purpose for restratification? Or is it also -- or is it -- just to make sure that -

DR. WEINTRAUB: Within a randomized trial, the reason for

stratification is to impose balance between groups. For reasons for -- you don't have to stratify this at all. You can still look at it as a subgroup. You could do that.

My reason I would think of these as separate strata is the description of resource use in a gamish that includes both PCI and CABG, doesn't to me make a lot of sense.

MS. TURBYVILLE: I just bring it up because some of the approaches that I've seen across developers is they'll estimate the measures as a whole, but then encourage the users to stratify to increase not -- to look at the whole, but also to increase actionability at a sub-population level. But I'm not trying to disagree, I'm just trying to think through --

CO-CHAIR CURTIS: But I think Bill's point is that those are subgroups as opposed to stratified analyses. Is that --

DR. WEINTRAUB: I mean, you should

always -- so how do you go about developing subgroups? The idea of subgroups in any analytic framework you want is for there to be a small number, have a pathophyisiologic basis and specify in advance. I'm not hearing a lot about proper development of subgroups in the measures we're talking about here.

MS. TURBYVILLE: I just wonder how much in -- that even NQF has indicated the difference between a strata and a subgroup, in thinking about how you propose users use the measure, not in any way disagreeing with what you're saying. I'm just wondering how much guidance we've actually given on that.

CO-CHAIR CURTIS: I think this is actually a minor point in the overall evaluation. But it does indicate some concern about the clinical sensibility approach, and particularly if you're going to stratify based on an outcome. This just doesn't make sense.

DR. WEINTRAUB: Yeah. Well, certainly stratifying based on outcomes makes

no sense.

CO-CHAIR CURTIS: We'll put that in the back parking lot and go on.

DR. WEINTRAUB: I mean, the whole thing where strata and subgroups comes from the clinical trial world, and you'd be applying it here. But I think there -- you know, there are reasons to think about that. Because that would overlap in their logic for reasons I raised.

Okay. So moving on. So then a lot of this is just sort of mechanical, on page 15, on how they're doing this.

And then they're using -- as I understand it, they're using the same approach to standardized pricing we heard about this morning. I didn't really fully understand it from their description here, but they explained it this morning.

Is your approach to pricing in this measure the same as in the AMI measure?

DR. STROUPE: Yes, it is. The

pricing approach is the same.

DR. WEINTRAUB: All right. I'm on page 16. And again, here they do mention the strata again here. Then we go to the next, is also on page 16 --

DR. HWONG: Sorry, could you go back? I know this morning when we talked about the Ingenix measure, it kind of went by a little fast.

DR. WEINTRAUB: Do you mean Ingenix or do you mean ABMS?

DR. HWONG: The --

DR. WEINTRAUB: You mean this same developer?

DR. HWONG: Yes. You know, previously I think it was taken off the table.

DR. WEINTRAUB: That was the Ingenix.

DR. HWONG: Right. Then me just sort of catching up and understanding the pricing again?

DR. WEINTRAUB: This is -- we

never even got a description of the pricing from them.

DR. HWONG: Okay.

DR. WEINTRAUB: This was the pricing from the MI measure we heard this morning from --

DR. HWONG: Okay.

DR. WEINTRAUB: -- what's this group, ABM.

DR. HWONG: ABMS, okay.

DR. WEINTRAUB: So they're using the same standardized pricing, based on payments.

You want to take -- very, very briefly take us through the costing strategy for the developer, very briefly?

DR. STROUPE: The costing strategies, there were three different methods for costing, depending on the type of utilization that we were using.

The -- for inpatient events, for inpatient and for -- for inpatient, there's

one for ambulatory, pharmacy, and then for all other events.

For the inpatient, a DRG was determined, and then using a pricing table for a cost per DRG, the length of stay that the patient incurred during the measurement period. So if the hospitalization occurred but then extended beyond the 12-month measurement period, those length of stay days weren't included. But for the length of stay days that were in the measurement period, that was multiplied by a cost for DRG to get the inpatient amount.

CO-CHAIR CURTIS: And this is overlapping or comparable to the NCQA methodology for defining inpatient, correct?

DR. STROUPE: That's correct.

And then there's some -- another coding in -- for situations without the DRG. So that's sort of a cost per day approach, was the approach that was used for the inpatient cost.

For the medication cost, for the data set that was available for the testing purpose, we took -- determined the -- for each NDC code, the base supply all the medication. And in addition to that, the cost for -- so for each NDC, the cost and the base supply.

And then determined for each NDC what an average cost per day would be, for each NDC code. And then for the data that -- for the patients that were in the measure, we would have information on the NDC code for their medication they received as well as the days supplied. So for each NDC we would multiply the days supplied by the cost per days supplied for that NDC code.

For the other, the ambulatory care, E&M, so forth, the -- a cost per the CPT code and CPT code modifier combination was determined for the entire data set. And then that cost was used to estimate a cost for each of the code events for the -- that occurred during the measurement period for the

individuals in the -- that were identified for the measure.

So those are the three approaches that were used to estimate the cost of inpatient, ambulatory, pharmacy and the other health care.

MS. CLARK: And the cost for CPT code modifier combination, that's coming from the Thomson-Reuters database?

DR. STROUPE: That's correct. That's by taking the data that were available and then using that to generate those values.

MS. CLARK: And is that including -- you know, if there was a facility component and a professional component? So for example, if somebody had a re-PCI and it was an outpatient procedure, which about 25 percent are outpatient, you're including the facility cost as well as the physician cost at the CPT modifier level?

DR. STROUPE: That would be for the -- it would be the cost that was in the --

within the database. And so that would be primarily for the facility component.

DR. WEINTRAUB: So how are you handling professional costing?

DR. STROUPE: The cost for the E&M, for things related to that. But basically regarding the cost for the CPT codes that were in the database.

DR. WEINTRAUB: All right. So you know, I think we have to -- one of our comments back to them that they'll have to provide better clarification of their costing strategies. I suspect that's true of the AMI measure as well, we didn't go through it in the detail we probably should have. Because we didn't spend enough time on it this morning.

(Laughter.)

DR. WEINTRAUB: All right. So I'm on page 16 to 17 of resource use service categories. And those are pretty good. They've already said that above so it sort of

becomes repetitious. And then they go identify this.

And then care settings, and that's all in the bottom of 17, it's pretty straightforward.

And then the risk adjustment methodology, as far as I can tell, is exactly the same as we heard this morning. So there's no perfect way of doing this, but it's as reasonable as any.

DR. STROUPE: That's correct. It was the same, the risk adjustment approach was the same as the other measure that was discussed this morning.

CO-CHAIR CURTIS: And again, just like that measure, I think there are some, given the 48 different models that were considered before deciding on which one was optimal, we need just more information as to the selection criteria for the final model, as well as the results for that model.

DR. STROUPE: That certainly seems

reasonable, based on our information.

DR. WEINTRAUB: And they have a model developed here, which is on page 19. So it's very difficult to really look at it critically and know, you can sort of specify it. And you can accept it or not accept it.

All right. So I'm at the bottom of 19 now, and they just take us through the their -- how they develop a model in some detail. So I would assume again it's essentially the same as for the AMI.

And now I'm on to page 20, S.10.2, stratification method. Again here we have this problem of the stratification by revascularization and follow-up.

Then we go through the costing methodology again, which we've just really discussed, going from page 20 to 21.

And I don't think we have to discuss that in detail. We want a little more -- a little better specificity from them. It's in detail here, but it's still kind hard

to figure out what they did in handling certain things, especially related to professional costing.

Okay, I'm not to -- now I'm going through all the costing. Ambulatory and pharmacy is on page 23.

CO-CHAIR CURTIS: I think we've covered pretty much everything up until the hierarchy, which is 25 -- or the attribution, sorry.

MR. AMIN: May I have a quick clarifying question, Bill?

DR. WEINTRAUB: Yes.

MR. AMIN: Was more -- did you need more robust information on pricing, standardized pricing for all or just the professional services?

DR. WEINTRAUB: So I didn't hear about any other problems, other than that. You know, the problems were sort of accepting what they're doing without fully understanding it, but I don't know what we could do about

that realistically.

MR. AMIN: Okay.

MS. TURBYVILLE: We could ask for more detail.

DR. WEINTRAUB: They provide a lot of detail. I think what we'd want is perhaps a little more -- some clarity, so maybe some overarching text of their approach to costing, and a little bit more of the professional component. Does that make sense?

All right, their approach to -- I'm going to keep going. Their approach to attribution, as far as I can tell is also the same as we heard with MI this morning. We're going to have the same kinds of problems, so I think that just at they need to address problems with attribution for MI, they're going to need to address problems with attribution for this measure as well.

All right. So if we come down to level of analysis, I think there's a problem. They're proposing a level of analysis at

individual clinician. I don't believe that that's going to be possible. I think measures like this will work at the level of health systems or health plans, but it would be very, very difficult to make this meaningful at the level of individual clinician.

So I think that should be part of our feedback to them, that they should rethink the level of analysis.

They're not providing specifications and guidelines for sample size requirements. That can be done, per the discussion this morning of doing some simulation work, of what it would take. It would be a help if they had a sufficient level of data to do that, but they don't right now.

So then this becomes more mechanical again through page 27, and the interpretation is it's standard O to E ratio sort of thing.

And I'm on to page 28 -- we're almost done, actually, with it. And will

provide reports along the lines we've heard, on the bottom of page 28.

So they've done some testing. Do we want to rate up to this point, and then we can very briefly go over their slides on testing? That will be faster than what we've just been through.

MS. TURBYVILLE: So you want to rate 2.A.1 and then after that move into the reliability and validity testing?

DR. WEINTRAUB: Yeah, we can do that. That's reasonable.

MS. TURBYVILLE: Okay.

CO-CHAIR CURTIS: So 2.A.1 and 2.B.1 I think are kind of like -- is that -- they're kind of there, right? Okay.

So regarding 2.A.1, the measure is well defined and precisely specified so that it can be implemented consistently within and across organizations.

DR. WEINTRAUB: So I think I rated it moderately, but we've come up with a lot of

problems. So I'm going to -- I'm actually going to rate it low. I think that this can be pulled up to snuff, but it will still require some reworking. Right now it's not ready.

But it's up to you.

CO-CHAIR CURTIS: I think -- I mean, specifically the exclusion of the MI population, I mean, to me that's strange and worthy of low. I mean, it makes it almost --

DR. WEINTRAUB: I think so.

CO-CHAIR CURTIS: -- uninterpretable as a measure.

DR. WEINTRAUB: Right. Also things like the stratification of something that occurs and follow-up doesn't make any sense. Their problems -- their costing methodology needs to be cleaned up a bit. There's a lot of problems here.

CO-CHAIR CURTIS: But the voting is in your hands.

So then for 2.B.1, the measure

specifications are consistent with the evidence presented to support the focus of measurement under criteria 1.B. It's specified to capture the most inclusive population. Bill do you want to comment or --

DR. WEINTRAUB: So we see again some of the problems in the population. I haven't really defined the problem of the target population properly and it needs a bit of work. I mean it's very global, it certainly can be fixable, but they're not there. So I think I rated it moderate before, but we've uncovered problems, and I'm going to rate it low again.

CO-CHAIR CURTIS: Eight lows. So I think this will require a lot of feedback, but it -- and overlapping -- I think you've captured it. Okay.

DR. WEINTRAUB: All right. So on page 29, so again, they're using the Thomson-Reuters market scan data set.

MS. CLARK: I wonder if we can

even rate these, the reliability and validity testing is --

MS. TURBYVILLE: They -- well, depending on the extent to it. But it is acceptable to do reliability and viability testing in a database.

MS. CLARK: No, but I mean if the measure is not even --

MS. TURBYVILLE: Oh, that's right.

MS. CLARK: -- if the measure is not even valid. We all rated it low, how can we test -- oh, they did a good job of testing it, but it's --

MS. TURBYVILLE: I think that's a good question. But well, I think that's part of how your rating will be impacted. So I mean, you'll - you can't help but have that influence it because that is a huge part of it. So staff will just make sure it's reflected in the notes that whatever rating you give is, in part, based on that if you'd like.

DR. WEINTRAUB: So they go to some effort to just describe the market scan database. And then they go into their analytic methods. And then they have the testing results. And that's on page 30.

The slides are actually helpful, and they begin on -- the computer is very colorful.

CO-CHAIR CURTIS: You can use the paper copy if you need.

DR. WEINTRAUB: Well, keeping everybody else on track, it will be helpful to get the slides.

So Jeptha took you through the slides before, and they're very similar. You can go slide 4 shows how they defined the population, and ultimately come down to 11,000 patients. And then they have their stratification -- I don't have that blue binding before, this is all over the place.

CO-CHAIR CURTIS: I will say that it's interesting that only 3.9 percent of

population is excluded, based on having had that MI between 14 and 365. I guess that's believable. It seems a little on the low side, but --

DR. WEINTRAUB: It seems low to me. I don't believe it. Maybe that suggests problems with the reliability of the data.

They go to, from there, it's slide 6. You see the same problem we saw this morning with the distribution of the costs, inpatient facility charges. The 75th percentile was at zero, and the 95th percentile was at 10,000. Very skewed data making it rather difficult to work with. Not impossible, but just the challenge needs to be understood.

And then in slide 7, we have the distribution of the costing. Slide 7 is the top 20 CAD-related E&M post-revascularization episodes. Just 14 percent of total episode costs. So that's pretty low total, low percentage.

And then they have a whole bunch of non-CAD related costs. Then they have -- then on the top of 9, the CAD-related procedures. And this is a big -- or a relatively big percentage of costs. That's pretty believable, actually.

CO-CHAIR CURTIS: Sorry, what slide?

DR. WEINTRAUB: Slide 10 is common non-related procedures. Well, you know, some of this can be argued about whether they're related or non-related. And this gets to the whole problem, again, if you're going to even try and go through this exercise of what's related and what's not. Replacement of aortic valve, cardiopulmonary bypass. Replacement of the mitral valve in particular I would not say is non-related with coronary disease.

CO-CHAIR CURTIS: And then specifically, like what's that $5.8 million of anesthesia for procedures on the heart pericardial sac and great vessels. That would

certainly seems to be relevant to the cost methodology.

DR. WEINTRAUB: And then insertion of a Swan-Genz catheter, well, your patients with coronary disease develop heart failure. That's why I think, you know, it becomes extremely difficult, different people are going to come up with different things they think they are related or not.

And the next thing is --

DR. LEE: This is Todd Lee, I just joined the call again. Can I clarify some of the information here?

DR. WEINTRAUB: Sure.

DR. LEE: For related and non-related information in that slide, for example the Swan-Genz catheter placement, 93-503, 2,418 of the times that occurred in our data, it was grouped to the episode. Only 13 times was not grouped to the episode. So while the primary grouping algorithm was by the diagnostic code or the procedure code, these

procedure codes were not specific to this episode.

And so the majority of these still are being counted in the episodes. So those $800,000 are being counted in the total episode costs here.

DR. WEINTRAUB: Oh, I see, 2,400 are related. Oh, you have it listed under non-related, but you have those related and not related all on the slide?

DR. LEE: Yeah, I apologize for that confusion. It's what we do with the intent of the slide was to say, look, sometimes if these procedures are showing up both as related and non-related, and we had asked our work group if they felt there was enough specificity with the procedure code itself to group it to the episode. If not, we still relied on the ICD-9 code.

And so this was our -- part of our iterative process with each of our work groups, to go through these codes. And they

would let us know, okay, yeah, now we're going to take that Swan-Genz catheter CPT code and group it to the episode, or no, we're comfortable with the proportion that seem to be grouping to the episode, let's just leave it with ICD-9 approach.

DR. WEINTRAUB: So I don't understand quite how you do that. If you put in a Swan-Genz catheter, how could it be -- and it's within the year's timeframe, how would you know it's not related?

DR. LEE: It was based on the ICD-9 code that showed up on that claim. So if the ICD-9 code that was with that specific CPT code for that claim, was not one grouped to the episode, it falls into this non-related column.

DR. WEINTRAUB: Well, okay. I think that's pretty hard to do, pretty hard to do it in a way that will be uniform for people.

CO-CHAIR ROSENZWEIG: Who does the

groupings? I'm not sure I understand who does the grouping to the episode?

DR. LEE: What process did we use for the grouping?

CO-CHAIR ROSENZWEIG: Yeah.

DR. LEE: This was -- each of our clinical advisory work groups, which is again principally made up of physicians, walked through each of these outputs with us. They generated an initial set of specifications that said, these codes are related to CAD, these codes should group to the episode.

And then as part of that, we would go through this output with them and they would identify additional codes that we would add to the specification, or codes that we might take away from the specification.

CO-CHAIR ROSENZWEIG: But with respect to like anesthesia for the procedures on the heart, how did they determine that there were 24 episodes that were not related to the event?

DR. LEE: There are 24 claims. These are 24 unique claims that do not have an ICD-9 code that is included in the ICD-9 code list for this episode.

CO-CHAIR ROSENZWEIG: But that's just a coding error.

DR. WEINTRAUB: Well, not necessarily. I mean --

DR. LEE: Or it's a procedure that's not related to their CAD. I mean, that's the other alternative. We don't know which direction -- we don't know if it's accurate or if there is some degree of misclassification. I mean, there is certainly potential for some noise in these measures, but this was part of the process that we worked through our -- with our clinical work groups.

CO-CHAIR ROSENZWEIG: Okay.

DR. WEINTRAUB: Well, I think this is very hard to pull off in a way that's going to be consistent and really believable. And

you're right, they may have come up with codes where it doesn't make sense. They're putting in a Swan-Genz catheter for patients in septic shock, I suppose is a possibility. But anesthesia for procedures of the heart, pericardial sac and great vessel, very hard to say that that's not related.

CO-CHAIR CURTIS: So maybe part of the feedback to you, the measure developer, would be that in trying to help us validate this approach by doing a deeper dive on some of these instances where there is uncertainty as to whether or not it's an appropriate classification. So digging in on those 24 claims for the cardiac anesthesia or what have you, some of the ones that are even more frequent, or the ones we identified earlier today.

I just think it would be -- it would go a long way to allaying sort of the suspicions as to this approach, which I grant you is reasonable and has some face validity

to it. But you know, it's just a matter of getting us comfortable with it as well, and maybe it's in all the outputs that the working groups saw over time, over the three years of the measure development. But it's just having the -- I think it's just hitting the wrong note with the group here, potentially.

DR. WEINTRAUB: I think you'll have to think about that for the heart failure measure we saw this morning as well. I mean, you have others here that are related sometimes where you'd think they'd probably never be related, if you're taking this approach.

Colonoscopy, flexible, proximal to splenic flexture. Fifty-eight of them are related. Cataract removal, 30 are related. So I mean, I'm not -- I'm having a fair amount of trouble here.

All right. So and then the next thing is imaging, CAD-related, and then non-related. And again, we have the same sort of

problem here on the imaging. Computed tomography, pelvis with contrast material, 186 are related, 365 are not. So I guess it's possible, you know, if they have a retroperitoneal hematoma. Pretty tough to pull off. You've got screening mammographies related in 231.

And then they have the testing on -- that's on slide 13. And then we have more testing. You have lipid panels, on page -- on slide 14, lipid panels, half are related and half aren't related.

And they have, on slide 15, major joint replacement, you have some of them that are related. And non-related -- okay.

All right, I think that's enough of this. I mean, I think there's some problems on how you're attributing what's related and not related. I think we've made that clear.

So why don't we go back to page 29, which is where we were before we started

all this. Jeptha, how do I get there rapidly?

Everyone should have Jeptha command their computer for them.

(Laughter.)

DR. WEINTRAUB: All right. So really, we're on page 30 with validity testing.

Okay. Are we ready to score then?

(Laughter.)

DR. WEINTRAUB: That was cute, Mary Ann.

MS. CLARK: Well, I was told.

DR. WEINTRAUB: All right.

CO-CHAIR CURTIS: All right, so we still have to do 2.A.2, reliability testing demonstrates results are repeatable producing the same results a high proportion of the time when it's in the same population in the same period.

Bill, what do you think?

CO-CHAIR ROSENZWEIG: Get your microphone.

CO-CHAIR CURTIS: I'm sorry.

CO-CHAIR ROSENZWEIG: Yeah, it wasn't me this time.

CO-CHAIR CURTIS: So to say it again, reliability testing demonstrates the results are repeatable producing the same results a high proportion of the time in the same population.

DR. WEINTRAUB: So we don't really quite buy into what they're doing here, so I don't think we have a -- they may measure at the same time, but I really thing that they have some real analytic problems here.

So I think I measured it as moderate, but with the problems we've uncovered, I'm going to rate this as low.

CO-CHAIR CURTIS: But I don't think this has to do with the scientific acceptability of it. This is a different criteria, which is just, if they apply this methodology in the same -- you know, how often can they get the same results, irrespective of

whether or not they're the right results, we know it's the same.

DR. WEINTRAUB: All right, so we'd have to say we actually don't know. Then we really have insufficient data. Because they have -- they've done this once. They haven't shown the kind of testing that it would seem -- that we just saw in the previous measure where it's been in use.

CO-CHAIR CURTIS: I think that's fair, but then again, being alert to the possibility of drift, and I think we're sort -- for the earlier one using the same methodology, I think I proposed at least a moderate, based on the fact that I think that they've gone through these codes in some detail and have at least a way that they're defining the costs the same way. And so I'll push on that -- I'm going to vote a little differently, but --

DR. REEDER: My light is on. They can be consistent regardless of what the

validity, what we determine construct validity to be. I think Dr. Curtis is correct in that we're looking at that consistency component. And it's a necessary component for the establishment of validity, but not all-inclusive or sufficient in and of itself.

So the consistency piece of it, I agree, is quite high here. And to be consistent across the measures or -- yeah, measures that we've done today, I would have to go with high.

DR. WEINTRAUB: High, that's interesting.

So but look at what it says here. Reliability testing demonstrates that the measure results are repeatable. We don't have a demonstration of that. It's been tested once.

Producing the same results a high proportion of the time. We don't know that.

CO-CHAIR CURTIS: I guess we don't know that, but I guess --

DR. WEINTRAUB: I mean, it's reasonable that it would. It looks like it would do the same thing again. But you don't know that.

MS. TURBYVILLE: But they did numerous -- just as a reminder, they did do numerous iterations of just removing 1-2 codes, and then -- so there was kind of repeat, repeat within what they submitted, as they submitted it.

CO-CHAIR CURTIS: So not to belabor it, I'd propose we just vote on this. So 2.A.2.

DR. WEINTRAUB: Wow, you can go anywhere you want on this one.

CO-CHAIR CURTIS: So one high, five moderate and two insufficient.

Regarding 2.B.2 which is validity testing, which I think is getting more into how it interplays with scientific acceptability, demonstrating that the measure data elements are correct and that the measure

score correctly cost of care or resources provided adequately distinguishing higher or lower costs for resource use.

Bill what --

DR. WEINTRAUB: I rated it moderate before. I think here we have some data. I'm going to rate it low.

CO-CHAIR CURTIS: Okay. So let's put that to the vote then.

Eight low.

2.B.3, exclusions are supported by the clinical evidence, otherwise they are supported by evidence of sufficient frequency of occurrence so that results are distorted without -- with the exclusion. And the other criteria. But I guess we're just focusing on the reasonability of the exclusion criteria.

DR. WEINTRAUB: So here the thing -- the problem is the MI's and how they handle it. I mean, it's really quite fixable on what they do, we'll have to do that.

CO-CHAIR CURTIS: So put that to

the vote.

I think we're getting better at voting, or pressing.

Two moderate, five low and one insufficient.

Criteria 2.B.4 we're at, correct? For outcomes measures or resource use measures, evidence-based risk adjustment strategy as specified and based on clinical factors influencing income -- or sorry, outcome --

(Laughter.)

CO-CHAIR CURTIS: -- and are present at the start of care.

And Bill, what did you think of this?

DR. WEINTRAUB: I'm leaving it here.

CO-CHAIR CURTIS: So I think it's a risk adjustment approach. And that includes the stratification.

DR. WEINTRAUB: Yeah, the

stratification makes no sense. And so -- otherwise, their approach for risk adjustment is fairly standardized. So they need to do a little bit of work here, clearly.

CO-CHAIR CURTIS: Okay. So let's vote on that.

So five moderate and three low.

2.B.5, data analysis demonstrates methods for scoring and analysis of the specified measure, allowed for identification of the statistically significant and practically/clinically meaningful differences in performance.

Bill?

DR. WEINTRAUB: So they haven't really demonstrated this yet. It's liable to work if they can fix their other problems.

CO-CHAIR CURTIS: Okay. Any other comments?

DR. WEINTRAUB: So I mean, I think right now, to me, it looks insufficient.

CO-CHAIR CURTIS: Okay, let's vote

on that.

One moderate, one low and six insufficient.

DR. WEINTRAUB: Well, I don't have that much influence with my family.

CO-CHAIR CURTIS: 2.B.6, multiple data sources; I think this is not applicable, so we won't vote on 2.B.6.

2.C, disparities have been identified, I think we can forego voting on that.

And that brings us to usability.

DR. WEINTRAUB: All right. So

CO-CHAIR CURTIS: Can you just remind us generally how we voted on usability for the related measure?

MS. WILBON: I'm actually trying to create a table side by side of the evaluations you did this morning on 1570 -- was it, no. 1571 and this measure so you can kind of see. I can bring that up in a second, but for now, let's see, for usability for the

measure this morning, right?

CO-CHAIR CURTIS: Right.

MS. WILBON: 3.A was rated eight insufficient -- yeah, they were all insufficient.

CO-CHAIR CURTIS: I just wanted to remind people that -- in general for this. And I think we can probably forego the review of it since we kind of know what we're going to say in this case, that they're similar.

So why don't we go ahead and vote on 3.A.

DR. WEINTRAUB: So when I originally rated these low throughout, but I think insufficient is a better descriptor.

CO-CHAIR CURTIS: And 3.B -- I guess it needs to tabulate, sorry.

So 3.B, meaningful results, go ahead and vote on that.

DR. WEINTRAUB: They're not there yet.

Waiting on one.

And 3.C, data results details.

DR. WEINTRAUB: Well, I want to see the results here first.

CO-CHAIR CURTIS: Okay, eight insufficient.

And moving to 3.C, whether or not it can be decomposed to facility transparency and understanding.

DR. WEINTRAUB: We can't tell about it yet, so again I would vote insufficient.

CO-CHAIR CURTIS: And 3.D we'll forego.

And for feasibility, I don't think we have the same insufficient --

DR. WEINTRAUB: No, yeah, I mean, this is all electronic. This can be done, that is routinely collected. So it is -- for this one, it is feasible. That I can actually rate high. Remember, not to say anything about believable as it is, but that you can do it.

CO-CHAIR CURTIS: So again, susceptibility to inaccuracies, errors and unintended consequences, is that the only one that might be worth discussing?

DR. WEINTRAUB: That's not a question, though.

CO-CHAIR CURTIS: Well, I know, but I thought we'd go through all the feasibility ones before going through the vote.

DR. WEINTRAUB: Oh, all right. Okay.

CO-CHAIR CURTIS: But I think the first two, I think we can agree on without much review. But 4.C, this morning I think we voted that it was --

MS. WILBON: One high, two medium and five low.

CO-CHAIR CURTIS: Yeah.

MS. WILBON: And then the last, 4.D was five moderate and three low.

CO-CHAIR CURTIS: Okay. So

starting the voting for 4.A, required data elements are routinely generated.

Six high, two moderate.

4.B, data elements are available in the electronic health record or other electronic sources.

DR. WEINTRAUB: Well, we don't know that.

CO-CHAIR CURTIS: That's administrative data. I don't know if it's different than any of the other measures that we --

DR. WEINTRAUB: Well, but this is -- I mean, what do the words say here? It says, are available in electronic health records or other data -- or other electronic sources, okay.

CO-CHAIR CURTIS: 4.C, susceptibility to -- I'm sorry. Six high, two moderate.

Susceptibility to inaccuracies, errors or unintended consequences.

DR. WEINTRAUB: All right. So here they've really got some clear-cut problems. So I'm going to vote low on this.

CO-CHAIR CURTIS: And 4.D, data collection measurement strategy implemented as demonstrated by operational use or testing do not identify barriers to operational use.

DR. WEINTRAUB: Well, until they do some more work, they've clearly got barriers. Of course, it's overcomable, but they've got some barriers as it is right now. So I'm going to vote low.

CO-CHAIR CURTIS: One moderate, five low and two insufficient.

And I think the feedback on this would probably more in line with the problems with identified or concerns we raised on scientific acceptability being the major barrier, probably in implementation and feasibility.

DR. WEINTRAUB: I would agree.

CO-CHAIR CURTIS: Okay. So that

was fun.

Let's wrap up the day. Let's go to Sally.

MS. TURBYVILLE: So before everyone leaves and has a chance to forget about today --

(Laughter.)

MS. TURBYVILLE: -- all of the staff who are supporting this wanted to quickly take -- it doesn't even have to be a minute of your time. Is there any adjustments we could make for tomorrow to make it go smoother in your estimation? So we can think about those adjustments ahead of time. Anything come to mind immediately?

Not to put people on the spot, but --

DR. MARWICK: I don't know about tomorrow, but from the one that I presented, I just wonder if there's -- we should consider some process of selection of these before they come to this meeting? Because there's some

that -- you know, that particular one was not on target, and I think there's others that are -- have got fundamental problems.

MS. TURBYVILLE: And I think that's an excellent question, and I appreciate you bringing it up.

We are relying on you as the clinicians to tell us if they're not measuring what's intended. Staff really tried to not -- while we looked for the submissions to be complete and work with the developers to make sure they're submitting things as much as we can, where we asked, if we preemptively held things back without getting your input, we would be playing the role of, you know, potentially biased.

DR. MARWICK: Okay.

MS. TURBYVILLE: So I think we do try to make sure that incomplete ones, but we don't want to make a clinical judgment without your input and guidance, because you really are the first stop in that assessment.

DR. MARWICK: I'm not suggesting that at all. I'm just suggesting there should be a process of culling ones -- after their sent out to us, culling ones that --

CO-CHAIR CURTIS: You're talking about maybe something analogous to fast-tracking for grant submission?

MS. TURBYVILLE: Oh, I see. So when we send it out. And hopefully in the future we'll be sending them out in four weeks' time for the TAP, they can -- there can be an opportunity through preliminary assessments to reconsider whether or not it should be dealt with onsite in the in-person meeting. Is that what you're -- yeah, I think that's a good idea.

CO-CHAIR CURTIS: I think so. And specific to that one, I mean, it took hours and hours and hours for Tom and I to go through that submission, and then ten minutes into it, they withdrew. And so if we'd been able to go through that earlier --

MS. TURBYVILLE: I still think we need the clinical input of someone to go through --

CO-CHAIR CURTIS: I think, yeah, on the basis of the reviewer rather than before, potentially --

MS. TURBYVILLE: Oh rather than the primary? Yeah --

CO-CHAIR CURTIS: We can explore different options.

MS. TURBYVILLE: It's a challenge, but we can certainly think out of the box a little bit.

DR. BURSTIN: There may be ways we can explore it, because I think maybe it will be a handful, I'm assuming. And maybe it's work with the two chairs and make sure that if you see -- if the member sees something like that, we can have a conversation with the developer and figure it out beforehand, perhaps.

MS. TURBYVILLE: As much as

possible.

DR. MARWICK: I think we were lucky that the developer took us off the hook, really, because we could have ground on doing something that we all knew was futile.

MS. TURBYVILLE: Point well taken.

DR. HWONG: And the only other suggestion I would have, maybe for the future, too, in terms of, you know, the 2.A section that had this like -- you know, a large number of the microcriteria, I remember sort of going through it, I felt like some of those criteria, you'd sort of talk about up front, and yet it's sort of listed again.

Maybe if we could group them, right, so that you can kind of -- so if it's exclusion criteria or, you know, something -- I would have to go back and take a look at the categories much more.

But I sort of felt like, yeah, going through, sometimes you'd end up sort of going back and you'd already covered it. And

I don't know how much of that is just, you know, offshoots of conversations. But somehow there might be a way to kind of group them a little more tightly.

MS. TURBYVILLE: Absolutely. We'll look at how we might do that. It is very much right now a laundry list, so I think we can put a little more thought, especially having the benefit of listening to a full day of review, how we might better group them.

CO-CHAIR CURTIS: The only other thing, I think the table that Ashlie is putting together, which will kind of remind us within each developer, what our previous assertions have been would be useful.

MS. CLARK: I just have one other comment. I really liked the NCQA's measure in terms of how they incorporated the quality measures into it. I thought that was very helpful. And what I'm wondering about, though, is you know, comparing these different ratios of costs to norms and things like that

is fine. But really, trying to focus more on tying it to the outcomes, because you could spend, you know, a certain amount of money and get much better outcomes than, you know, maybe somebody who's spending less money.

So I guess I'm just wondering if we're going to make that -- or if anyone is going to make that next leap to tying the cost to outcomes.

CO-CHAIR CURTIS: We discussed this at length at the steering committee, and that we didn't want to have an unlevel playing field such that those measure developers who hadn't already pre-tied it to quality would be looked on more favorably. Intuitively you kind of feel that. But we thought that it wouldn't necessarily be fair, or we didn't want to suppress interest if they hadn't made that connection already.

So I think that was the compromise that we made.

DR. BURSTIN: We probably should -

- not that anyone is in the room, and I don't know if anyone is on the phone, but we probably should have public comment.

And then maybe start it out in the morning and see if there's anyone, since I have a feeling we've dropped off the -- Operator, can you see if there's anyone who has any public comments?

Operator, are you there?

All right, so I guess we'll start with public comment.

OPERATOR: Again, if you have a comment, star-one.

MS. TURBYVILLE: Anything else, Ashlie? Actually, let me make sure.

(This proceeding was concluded at 5:45 p.m.)

D

H

I

N

T

V

C E R T I F I C A T E

This is to certify that the foregoing transcript

In the matter of: Technical Advisory Panel

Before: NQF

Date: 05-10-11

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